

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representation of
The original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**

= GB 2336698

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : G06F 17/30		A1	(11) International Publication Number: WO 99/56223 (43) International Publication Date: 4 November 1999 (04.11.99)
(21) International Application Number: PCT/GB99/01213 (22) International Filing Date: 21 April 1999 (21.04.99) (30) Priority Data: 9808805.7 24 April 1998 (24.04.98) GB (71) Applicant: THE DIALOG CORPORATION PLC [GB/GB]; The Communications Building, 48 Leicester Square, London WC2H 7DB (GB). (72) Inventor: BAGSHAW, Marcus, Alexander; 69 Victoria Road, Chelmsford, Essex CM1 1PA (GB). (74) Agent: ATKINSON, Ralph; Atkinson & Co., The Technology Park, Shirland Lane, Sheffield S9 3PA (GB).			(81) Designated States: JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published With international search report.
(54) Title: ASSOCIATING FILES OF DATA			
<pre>graph LR 221[221] --> 222([DIVIDE INTO SECTIONS]) 221 --> 223[223] 222 --> 224([CATEGORISE SECTIONS]) 224 --> 225([PROCESS SETS OF ASSOCIATIONS]) 225 --> 226([ASSOCIATE]) 226 --> 227[(DATA BASE)] 223 --> 226</pre>			
(57) Abstract <p>Data files are associated with categories by processing said data files in combination with outline files. Large files (221) are divided into a plurality of file sections (223) each having a size substantially consistent with a preferred size. Each of the file sections is categorised (224) and the sets of associations are processed (225) to produce a set of category associations for the original undivided file (221).</p>			

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

Associating Files Of Data

Field of the Invention

5 The present invention relates to associating large data files to information categories.

Introduction to the Invention

10 Traditionally, database technology has been dedicated to the organisation of numerical and tabular data and it is only recently, particularly with the expansion of the Internet, that demand has grown significantly for the retrieval of text-based files. Several facilities are available on the Internet, commonly referred to as "search engines" which assist in the location of information. The majority of these operate by performing what has become known as "free text" searching, in which a user specifies words which they
15 believe are contained within the target file as a mechanism for instructing the system to retrieve files of interest.

Problems with this technique are well known to users of the available search engines, and a simple enquiry can generate hundreds of thousands of "hits", the majority of which will tend to be totally irrelevant to the user's
20 needs. Furthermore, other relevant files may be missed simply because they do not contain the specific chosen words.

As is well documented, a problem with the Internet is that the freedom of the Internet is also its downfall. Information is not classified before it is made available, therefore it is highly likely that even the simplest search will
25 fail to identify relevant documentation and will take a considerable period of time to perform.

Procedures for classifying volumes of data so as to facilitate

subsequent searching are known but these classification processes often involve manual intervention, thereby making them time consuming and prone to human error. Furthermore, except in circumstances where the documentation is considered to be extremely valuable and will continue to be required over a significant period of time, the cost of performing this manual exercise cannot be justified in terms of the commercial worth of the data sources being considered. Consequently, the problem results in much data being effectively inaccessible and outside the realm of searchable knowledge.

Procedures are known for processing a data file so as to determine whether the data file should be associated with a particular information category. The known processes require a machine readable association file (or outline file) and using this, it is possible for the incoming data file to be processed to produce a numerical score value defining the extent to which the data file is relevant to the associated category. Thereafter, decisions may be made as to whether the data file is to be associated with particular categories by performing respective threshold comparisons.

In practical systems, thousands of such outline files would be required in order to provide a useful level of categorisation. In the present applicant's co-pending British patent application number 98 08 808.1, in the present applicant's co-pending European patent application (DGC-P11-EP) and in the present Assignee's co-pending United States patent application (DGC-P11-US) a method of generating machine readable association files is described. A plurality of data files are manually selected as being examples of files which should be associated with a particular category. In addition, a plurality of files are selected manually which are considered not to be associated with a particular category. Having identified these files, the

process identifies preferred term candidates from the associated files, weights these candidates with reference to files not associated with the category and applies terms to a machine readable association file by analysing the weighting values.

5 The resulting association files are particularly well suited to associating new data files which are of substantially similar size to the original source data files. Similarly, association files generated by more traditional techniques still tend to be well suited to input data files of a particular size but less well suited to incoming data files of differing sizes. Thus, if a new incoming data
10 file is larger than the optimum file size, it is possible that many irrelevant files will be inappropriately categorised given that the processing of these files will result in an inappropriately high weighting value being calculated.

Summary of The Invention

15 According to a first aspect of the present invention, there is provided apparatus configured to associate files of data of a size greater than a predetermined size, comprising dividing means configured to divide a file into a plurality of file sections having a size substantially consistent with a preferred size; categorising means configured to categorise each of said file
20 sections to produce sets of section associations; and processing means configured to process said sets of section associations to produce a set of category associations for the original undivided file.

 In a preferred embodiment, the categorising means is configured to categorise each file section by processing a section in combination with
25 association files. Preferably, the apparatus includes storage means for storing the association files as outline files and each of the stored outline files may relate to a respective category.

According to a second aspect of the present invention, there is provided a method of associating files of data of a size greater than a predetermined size, comprising steps of dividing a file into a plurality of file sections each having a size substantially consistent with a preferred size; categorising each of said file sections to produce sets of section associations; and processing said sets of section associations to produce a set of category associations for the original undivided file.

Preferably, the preferred size is smaller than the predetermined size.

In a preferred embodiment, tables are removed from a data file before a file is divided into sections. Preferably, an assessment is made as to whether it is desirable to increase size sections, whereafter the size of said sections are increased and the dividing process is repeated. Preferably, data files are continually received from data sources.

Brief Description of The Drawings

Figure 1 shows a data distribution environment, including a data processing, storage and retrieval system;

Figure 2A illustrates procedures performed by the processing system shown in *Figure 1*, including a process for specifying preferred terms, a process for associating preferred terms with source files and a process for performing a search in response to a user request;

Figure 2B shows an overview of the operations performed by the environment shown in *Figure 1*;

Figure 3 details a processing system identified in *Figure 1*;

Figure 4 details a process for specifying preferred terms for association with data files identified in *Figure 2*;

Figure 5 details a process for the generation or modification of outline

files identified in *Figure 4*;

Figure 6 illustrates a graphical view of an OTL file structure;

Figure 7 illustrates a data file for the information represented graphically in *Figure 6*;

5 *Figure 8* shows a diagrammatic representation of the data file shown in *Figure 7*;

Figure 9 details a subsidiary processor of the type identified in *Figure 3*;

10 *Figure 10* details procedures for the creation of a new OTL file identified in *Figure 5*;

Figure 11 illustrates a plurality of rulebases produced by the execution of procedures identified in *Figure 10*;

Figure 12 details a process for the association of preferred terms identified in *Figure 2*;

15 *Figure 13* details the processing of a section to obtain a list of associated preferred terms identified in *Figure 12*;

Figure 14 details a triggering phase identified in *Figure 13*;

Figure 15 details a scoring phase identified in *Figure 13*;

Figure 16 details a list generation phase identified in *Figure 13*;

20 *Figure 17* details the processing of section results identified in *Figure 12*;

Figure 18 details a table of preferred terms;

Figure 19 details a linked list of preferred terms;

25 *Figure 20* details procedures for performing a search identified in *Figure 2*;

Figure 21 details a screen display prompting a user to identify a search method;

Figure 22 details a screen display prompting a user for search criteria; and

Figure 23 details a screen display for displaying titles of associated files.

5

Detailed Description of The Preferred Embodiments

A data distribution environment is illustrated in *Figure 1* in which data, received from a plurality of data sources 101, 102, 103 is supplied to a data processing, storage and retrieval system 104. Data sources 101 and 102 supply data directly to processing system 104 while data source 103 supplies data via a local area network 105, thereby allowing user terminals 106 and 107 to gain direct access to their local data source 103.

The processing system 104 provides access to a plurality of users, such as users 111, 112, 113, 114, 115, 116 and 117. User 111 has direct access to the processing system 104 while users 112, 113 and 114 gain access to the processing system 104 via the Internet 118. Users 115, 116 and 117 exist within a more sophisticated environment in which they have access, via a local area network 119 to their own local database system 120 in addition to a connection, via an interface 121, to the data processing system 104.

All incoming data from data sources 101 to 103 is categorised with a key word in seven separate fields, comprising "market sector", "location", "company name", "publisher", "publication date" and "scope". Users, such as users 112 to 117 may specify almost any term as the basis for a search and are then prompted by an equivalent word or phrase which constitutes more preferred search parameters. For example, a user may specify a search word such as "confectionery" and the system will prompt the user to consider

narrower terms such as "chocolate" along with related terms such as "cakes" or "desserts", or broader terms such as "food". From a simple request, a user is given an option of focusing further or of taking a broader overview of the subject under consideration.

5 The scope of an article refers to the context in which the document or article was written. For example, the scope field may consider questions as to whether the article concerns "mergers and acquisitions" or "seasonal trends" et cetera. Such terms are useful in gathering related information from a wide variety of industries and markets and may prove invaluable for particular
10 applications.

 The same criteria used for indexing are offered for search purposes and the same indexing terms are used for all documents across a range of specific databases. An overview of procedures performed by the data processing system 104 is illustrated in *Figure 2*. At step 201 preferred terms
15 for association with data files are specified. This step is essentially performed as an "off-line" process; establishing the environment for allowing source data to be processed as it is received from sources.

 Steps 202, 203 and 204 represent on-line procedures after the preferred terms have been specified at step 201. At step 202 the processing
20 system 104 receives data from sources such as sources 101, 102 and 103. The source data may be transmitted using different protocols, formats and standards therefore the processing system performs a standardisation process so that the data may be stored locally at the data processing system using standardised formats.

25 At step 203 the data is processed so as to enhance a user's ability to identify information of interest. Files of machine-readable data received from the sources are associated with specific preferred terms which

may be considered as defining particular information types. A file is considered and individual data elements, usually in the form of natural language words, are examined to identify occurrences of specified data types. The purpose of this association is to identify files of data which are of interest in relation to particular topics. This enables a user to organise a search which should result in useful information being supplied to said user, with reference to said topics and defined terms, from an extremely large database of stored data files. In this way, the technical procedures performed by association step 203 significantly enhances the overall functionality of the system and provides an industrially applicable approach to allowing highly focused sets of information to be supplied to a user in preference to large volumes of data; much of which will tend to be totally irrelevant.

In order to achieve this, the files of data are processed and are given a score representing a numerical value as to their relevance with respect to the predefined topics. Scores are adjusted in response to the number of identified occurrences of a specified data type. Furthermore, these scores are also adjusted in relation to the size of the data contained within the file. In particular, occurrences of data types in relatively small files are given a higher weighting with occurrences in larger files being given a lower weighting. Thus, the adjustment of scores is related inversely to the actual size of the data file. Thereafter, a threshold for the scoring values may be set and information types are associated with particular files dependent upon whether particular value scores fall on one side or on the other side of this threshold.

At step 204 a search is performed, in response to preferred terms identified by a user such that information of interest may be identified within the data stored by the data processing system 104 and transmitted to user terminals, such as terminal 111, over transmission channels as illustrated in

Figure 1.

An overview of the operations performed by the environment shown in *Figure 1*, in accordance with the present invention, is shown in *Figure 2B*. Processing system 104 receives input files, such as input file 221. It has been
5 determined that incoming file 221 has a size which is greater than a predetermined size and in this example the predetermined size is set at fifty thousand characters. A process 222 divides the incoming file into a plurality of file sections each having a size substantially consistent with a preferred size. In this example, the preferred size is established at ten thousand
10 characters. Furthermore, a file section is considered to have a size consistent with this preferred size if it has an actual size of between ten and twenty thousand characters.

File sections derived from file 221, in response to the operation of process 222, are illustrated generally at 223. Six file sections are shown but
15 the actual number of file sections produced will depend upon the size of the original file.

File sections 223 are each individually categorised by process 224 and it is assumed that a process is available for performing this categorisation upon files having sections consistent with the preferred size.

20 Process 224 produces a set of associations for each section and the sets of section associations are processed by process 225 to produce a single set of category associations for the original undivided file 221. An association process 226 then associates the original file 221 with the set of associations produced by process 225 so that the file data and its association
25 data may be written to a database 227.

Processing system 104 is detailed in *Figure 3*. Data signals from data sources 101 to 103 are supplied to input interfaces 301 via data input lines

302. Similarly, output data signals are supplied to users 111 to 117 via an output interface 303 and output wires 304. Input interface 301 and output interface 303 communicate with a central processing system 305 based on DEC Alpha integrated circuitry. The central processing system 305 also communicates with other processing systems in a distributed processing architecture. Processing system 104 includes eight Intel chip based processing systems 311 to 318, each implementing instructions under the control of a conventional operating system such as Windows NT.

An operator communicates with the processing system 104 by means of an operator terminal, having a visual display unit 321 and a manually operable keyboard 322. Data files received from sources 101 to 103 are written to bulk storage devices 323 in the form of large magnetic disk arrays. Data files are written to disk arrays 323 after these files have been associated with preferred terms, as illustrated at step 203. These association processes are performed by the subsidiary processors 311 to 318 and the central processing system 305 is mainly concerned with the switching and transferring of data between the interface circuits 301, 303 and the disk arrays 323.

The central processing system 305 communicates with the subsidiary processors 311 to 318 via an Ethernet connection 324 and processing requirements are distributed between processors 311 to 318. Having addressed a subsidiary processor 311 to 318 the transferring of data to an addressed processor is performed. Each individual incoming data file is supplied exclusively to one of the subsidiary processors. The selected subsidiary processor is then responsible for performing the association process, to identify preferred terms relevant to that particular data file. Thereafter, the associated data file is returned to the central processing

system 305, over connection 324 and the central processing system 305 is then responsible for writing the associated data file to the disk array 323. In this way, it is possible to scale the degree of processing capacity provided by system 104 in dependence upon the volume of data files to be processed in this way. The central processing system 305 also maintains a table of preferred terms, pointing to particular data files which have been identified as relevant to said preferred terms. The facility includes a CD ROM reader 325, arranged to read CD ROM's such as ROM 326. In this way it is possible to install executable instructions for computer system 305 and for computer systems 311 to 318.

Process 201 for specifying preferred terms for association with data files is detailed in *Figure 4*. At step 401 a preferred term is selected and at step 402 an outline (OTL) file is generated or modified. At step 403 a question is asked as to whether another term is to be processed and when answered in the affirmative control is returned to step 401, allowing the next term to be processed at step 402. Eventually, all of the terms will have been processed resulting in appropriate generations or modifications to their related outline files. Consequently, the question asked at step 403 is answered in the negative whereafter at step 404 data structures are initialised by parsing the OTL files generated at step 402.

Step 402 for the generation or modification of outline files is detailed in *Figure 5*. At step 501 a visual OTL editor is opened resulting in the editor's visual interface being displayed on VDU 321. At step 502 a question is asked as to whether an existing file is to be loaded for modification and if answered in the negative a new OTL file is created at step 503. If the question asked at step 502 is answered in the affirmative, step 503 is bypassed and at step 504 modifications or additions are made to the OTL definition. At step 505 the

OTL modifications created at step 504 are tested on a sample of test data and at step 506 a question is asked as to whether another modification is to be made. When answered in the affirmative, control is returned to step 504 resulting in further modifications or additions being made to the OTL definitions. When answered in the negative at step 506, the new or modified OTL file is saved at step 507.

When performing modifications or additions at step 504, a graphical representation of the OTL file data is presented to an operator via the visual display unit 321. An example of a display of this type is illustrated in *Figure 6*, representing a graphical illustration of a specific OTL file.

The OTL file stores definitions in an hierarchical tree structure and this structure is represented in the graphical view as shown in *Figure 6*. A representation of the tree may be contracted or expanded and the possibility of expanding a particular branch is identified by a plus sign on a particular line, as shown at 601. Similarly, when a particular branch has been fully expanded, the line is identified by a minus sign as shown at 602. Definitions within the file consist of rules, words and labels. The labels allow relationships to be defined between various parts of the file and between individual files themselves. The words identify specific words within an input file of interest and the rules define how and what weights are to be attributed to these words. Each rule line includes, at its beginning, a weight value 603 representing the score that will be attributed when a particular rule condition is met. Rules may also have leaves and the rule defines the way in which scores generated from leaves are combined.

OTL file data, represented graphically in the form shown in *Figure 6*, is actually stored in a data file having a format of the type shown in *Figure 7*. The actual data file shown in *Figure 7* corresponds to the data display in

Figure 6 but in Figure 7 all of the data, some of which has been rolled up in Figure 6, is present. The data contained within the file shown in Figure 7 is manipulated interactively by an operator in response to the graphical interface displayed as illustrated in Figure 6. Score values 603 are also
5 identified in the data file shown in Figure 7.

Displayed line 601 in Figure 6 is generated from line 701 of the actual stored data. The syntax of the language used for recording the data, as illustrated in Figure 7, may vary and the example shown is specific to this particular application. However, the underlying functionality of the language
10 may be considered with reference to the diagrammatic representation shown in Figure 8.

The outlines analyse data files in order to produce numerical evidence as to the relevance of a particular file with relation to a particular topic. The OTL definitions and structures are determined empirically and would be
15 modified and upgraded over a period of time. The system does more than merely register the existence of a particular word item by placing the word items within an interacting structure; the nature of which is illustrated in Figure 8. The particular entry, given label "oil-industry-mkt" relates to marketing aspects of the oil industry and as such can contribute to an overall
20 score as to the pertinence of incoming data to this particular topic. The first line 801 shows that this particular contribution may provide a total score of forty percent. This total of forty percent is then subdivided such that at line 802 the presence of the phase "buying oil from" has a score of fifty percent. Thus, the total contribution made by the presence of this phrase consists of
25 fifty percent of forty percent, i.e. a total of twenty percent being made to the total contribution. Similarly, as shown at line 803 and below, particular words may be identified which result in contributions of sixty percent of thirty percent

of forty percent. Thus, a complete OTL file is structured in this way with particular words and phrases making contributions to an overall score value. These words and phrases may also be specified in the rules as making single contributions or being allowed to accrue.

5 Examples of score value 603 are illustrated in *Figure 8* at 804 to 815. The hierarchical structure in *Figure 8* consists of a plurality of branches with lowest level entries being considered as leaves. Each leaf has a score value associated with it and values 809 to 815 are leaf score values. Above these, branch score values exist such that score values 807 and 808 exist at the
10 lowest level of branching with score values 805 and 806 being at the next level of branching further up each connected to the highest level of branching illustrated by score value 804.

 A total score value for a particular occurrence, detected at any level within the hierarchical tree, results in a final score contribution derived from
15 the product of the score value assigned to that particular level with all score levels identified while ascending the tree structure back to its root.

 The score contributions are produced and possibly accumulated when occurrences of the specified elements are identified. This provides a numerical weight to assess whether a file being processed should be
20 associated with a particular information type. The present invention, as implemented within the preferred embodiment, further adjusts these score weight in relation to the size of the overall data file. In particular, it is the branch score values (804 to 808) which are modified in preference to leaf values (809 to 815).

25 Subsidiary processor 311 is detailed in *Figure 9*. The processor includes an Intel Pentium processing unit 901 connected to sixty-four megabytes of randomly accessible memory 902 via a PCI bus 903. In

addition, a local disk drive 304 and interface circuit 905 is connected to bus 903. Interface circuit 905 communicates with the TCP/IP network 324.

Random access memory 902 stores instructions executable by the processing unit 901, in addition to storing input data files received from the data sources 101 to 103 and intermediate data. Procedures 503 for the
5 creation of a new OTL file are detailed in *Figure 10*.

At step 1001 temporary memory structures are cleared and at step 1002 an OTL description file is selected. At step 1003 an item in the OTL file is identified and at step 1004 a question is asked as to whether the
10 item selected at step 1003 is a rule definition. If this question is answered in the affirmative, a rule object is defined at step 1005. Alternatively, if the question asked at step 1004 is answered in the negative, to the effect that the item is not a rule definition, a question is asked at step 1006 as to whether the item is a word definition. If this question is answered in the
15 affirmative, a dictionary link is created at step 1004.

At step 1008 a question is asked as to whether the item is a label and when answered in the affirmative a new entry is created in a label list, whereafter at step 1010 a question is asked as to whether another item is present. After executing step 1005 or after executing step 1007, control is
20 directed to step 1010.

When a question asked at step 1010 is answered in the affirmative, to the effect that another item is present, control is returned to step 1003 and the next item is identified in the OTL file. Eventually, all of the items will have been identified resulting in the question asked at step 1010 being answered
25 in the negative. Thereafter, at step 1011 a question is asked as to whether another OTL file is present and when answered in the affirmative control is returned to step 1002 allowing the next OTL description file to be selected.

Thus, this process continues until all of the OTL files have been considered resulting in the question asked at step 1011 being answered in the negative.

For each OTL file considered, by being selected at step 1002, a rulebase is generated and a plurality of such rulebases is illustrated in *Figure* 11. Thus, a first OTL file processed in accordance with the procedures shown in *Figure 10* results in the generation of a first rulebase 1101. Similarly, further iterations of the procedures shown in *Figure 7* result in the generation of rulebases 1102 to 1109. Typically, for a specific installation, in the order of three thousand rulebases would be generated by execution of the procedures illustrated in *Figure 10*.

Rulebases 1101 to 1109 are stored in memory 902, which also provides storage space for a dictionary 1121, a label list 1122 and a data buffer 1123. The dictionary stores a list of words which have importance in any of the stored rulebases. Associated with each word in the dictionary, there is at least one pointer and possibly many pointers, to specific entries in specific rulebases 1101 to 1109. Thus, the words identified at 803 in *Figure 8* would all be included in dictionary 1121. Entries within the dictionary 1121 are implemented upon execution of step 1007 in *Figure 10*. Similarly, execution of step 1009, creating a new entry in the label list, allows a label to relate to rules that are elsewhere in the tree structure.

Process 203 for the association of preferred terms with source files is detailed in *Figure 12*. At step 1201 a question is asked as to whether the file is larger than a predetermined file size and if answered in the negative, control is directed to step 1209 where the whole file is processed to obtain a list of associated preferred terms.

If the question asked at step 1201 is answered in the affirmative, to the effect that the file is larger than a predetermined size, any tables present

within the file are removed at step 1202.

At step 1203, a preferred section size is selected and at step 1204 a file section is selected for processing and at step 1205 the section is processed so as to obtain a list of associated preferred terms. At step 1206 a
5 question is asked as to whether another section is present and when answered in the affirmative control is returned to step 1204 where the next section is selected.

Eventually, all of the file sections will have been processed and the question asked at step 1206 will be answered in the negative. At step 1207
10 section results are processed to select reliable preferred terms. After data association at step 1207, the data is stored at step 1210 with its associated preferred terms and data pointers associated with the preferred terms are updated at step 1211.

A plurality of possible techniques are available for dividing large files
15 into a plurality of sections. Firstly, the file could be divided on strict arithmetic grounds with section boundaries being equally spaced and derived purely on a character count. Thus, sections could be divided at two thousand characters or, more consistent with data storage environments, they could be divided at two kilo-bytes; equivalent to one thousand and twenty-four
20 characters.

Alternatively, the file may itself be structured with headings and sub-headings etc. Thus, it may be preferable to divide the file at the start of headings or, if this is not possible, at the start of sub-headings. Experience
25 has shown that the procedures are more effective if the file is divided at positions related to its actual content.

Procedures 1205 for the processing of sections to obtain lists of associated preferred terms are detailed in *Figure 13*. The overall processing

is broken down into three major phases, consisting of a triggering phase at **1301**, followed by a scoring phase at **1302** followed finally by a list generation phase at step **1303**.

Triggering phase **1301** is detailed in *Figure 14*. At step **1401** a section
5 of the data, such as its title, market sector or main body of text, is identified
and at step **1402** an item of the identified section is selected. At step **1403** a
question is asked as to whether the item indicates a new context, which may
be considered as a grammatical marker in the form of a full stop, capital, start
of a sentence or quotation marks et cetera. When answered in the affirmative
10 new context information is supplied to all rulebases **1101** to **1109** at step
1404 and control is then directed to step **1407**.

If the question asked at step **1403** is answered in the negative,
step **1404** is bypassed and a look-up address is obtained for rule objects in
rulebases from the dictionary at step **1405**. Thereafter, at step **1406** all
15 addressed objects are triggered and a multiplication of scores is effected by a
score weighting factor. Thereafter, at step **1407** a question is asked as to
whether another item is present and when answered in the affirmative control
is returned to step **1402**.

Eventually all of the items for a selected section will have been
20 considered resulting in the question asked at step **1407** being answered in
the negative. Thereafter, at step **1408** a question is asked as to whether
another section is to be considered and when answered in the affirmative
control is returned to step **1401**. At step **1401** the next section is identified
and steps **1402** to **1408** are repeated. Eventually, all of the sections will have
25 been considered and the question asked at step **1408** will be answered in the
negative.

As shown in *Figure 8* each lowest level leaf of the hierarchical tree has

a numerical value associated with the identification of a particular item, as identified generally at 803. If the amount of data contained within a particular file is less than what would generally be accepted, scores are further adjusted in relation to this size so as to improve the association of files with particular information types. This further adjustment is performed at the lowest leaf level (an example of this being level 803 in *Figure 8*) and these leaf values are multiplied by a file weighting factor, derived from the size of the file, which is triggered at step 1406 as shown in *Figure 14*.

Scoring phase 1302 is detailed in *Figure 15*. At step 1501 a rulebase is selected and at step 1502 a score variable is re-set to zero. At step 1503 a branch is identified for score accumulation/accrue and at step 1504 scores are accumulated or accrued from triggered rules attached to the branch. At step 1505 a question is asked as to whether another branch is to be considered and when answered in the affirmative control is returned to step 1503. A next branch is selected at step 1503 with procedure 1504 being repeated. Eventually all of the branches will have been considered resulting in the question asked at step 1505 being answered in the negative.

At step 1506 an overall score in the range of zero to one hundred is stored for the rulebase and at step 1507 a question is asked as to whether another rulebase is present. When answered in the affirmative control is returned to step 1501 and steps 1501 to 1507 are repeated. Eventually, all of the rulebases will have been considered and the question asked at step 1507 will be answered in the negative.

Phase 1303 for the generation of a list of associated preferred terms is detailed in *Figure 16*. At step 1601 a rulebase is identified having a score greater than a predetermined threshold. Thus, for a particular application a threshold may be set at forty-eight percent. At step 1602 additional triggered

preferred data characteristics are identified by associating successful rulebases with parent categorisations by rulebase links.

At step 1603 lists of successful and inferred rulebases are combined to form overall lists of preferred data characteristics. Step 1603 results in data
5 being generated by a subsidiary processor, such as processor 311, which is then supplied back to the central processing system 305 over interface 325.

Process 1207 for the processing of section results to select reliable preferred terms is detailed in *Figure 17*. At step 1701 sections with no associated preferred terms are removed from the scoring process, that is to
10 say they are ignored and at step 1702 a variable N is set equal to the number of remaining sections.

At step 1703 the number of occurrences for each preferred term is counted for the file as a whole, that is to say, individual counts for each set of occurrences are combined. At step 1704 a percentage of occurrences of
15 preferred terms is calculated with respect to N, as calculated at step 1702. Thereafter, at step 1705 triggered preferred terms are removed if their percentage occurrence falls below a threshold value, defined in terms of the percentage number of times a category should be triggered.

A category could be triggered by each of the sections therefore the
20 total number of possible triggers is equivalent to the number N of remaining sections. Thus, a percentage occurrence value is given by the number of sections which did trigger a particular category divided by the total number of sections then multiplied by one hundred.

The purpose of step 1705 is to remove associations that may be
25 considered as mistakes. Such associations are identified as mistakes if categories are triggered by only relatively few of the individual sections. Process 1701 and process 1705 both remove associations to preferred terms

and it is possible that too many associations may be have been removed, a situation that is likely to occur if the section size selected is too small.

At step 1708 the average occurrence of the remaining preferred terms is calculated and at step 1709 preferred terms scoring above the average
5 calculated at step 1708 are selected as being reliable for association with the original large file. As an alternative to rejecting associations falling below the average at step 1709, such associations may be retained as possibly reliable associations in addition to the reliable associations.

Referring to *Figure 18*, the preferred term "OIL_INDUSTRY" is shown
10 in first column 1801 associated to a pointer 0F8912 in column 1802. Address 0F8912 is the first in column 1901 of a linked list shown in *Figure 19*. Column 1902 identifies a particular file name and column 1903 identifies the next pointer in the list. Thus, entry 0F8912 points to a particular file with the file name "OIL_INDUSTRY_NETHERLAND_3" with a further pointer to memory
15 location 0F8A20. At memory location 0F8A20 a new file name is provided, illustrated at column 1902 and again a new pointer is present at column 1903. Eventually, all relevant files will have been considered and the end of the list is identified by address 000000 at the pointer location in column 1903.

In an active system, the database 323 will be continually updated and
20 users will continually be given access to the database, all under the control of the central processing system 305. Thus, with reference to *Figure 2*, it should be understood that the association step illustrated at 203 and the searching step at 204 are actually concurrent and will be effected in response to the availability of data and the demand for searching respectively.

25 Procedures 204 for performing a search in response to a user request are detailed in *Figure 20*. At step 2001 a user logs onto the system and at step 2002 a search method is identified. At step 2003 search criteria are

defined and at step 2004 search criteria are processed to determine preferred terms. At step 2005 a list of preferred terms are supplied to the central processing system 305.

5 At step 2006 a question is asked as to whether the host has responded and when answered in the affirmative titles of associated data files are displayed at step 2007.

At step 2008 a question is asked as to whether the user wishes to view identified data and when answered in the affirmative the data is viewed; after being downloaded over the communication channel, at step 2009.

10 At step 2010 a question is asked as to whether another search is to be performed and when answered in the affirmative control is returned to step 2002.

Step 2002 requires the search method to be identified and in order to achieve this a user is prompted by a screen display of the type shown in
15 *Figure 21*. Thus, a plurality of text boxes are presented to the user inviting the user to specify a search method.

Step 2003 for the defining of search criteria results in the user being prompted by a screen of the type shown in *Figure 22*. Terms providing a basis for the user's search are displayed in a window 2201. Preferred terms
20 are displayed in uppercase characters, such as the entry shown at position 2202.

The displaying of titles of associated files at step 2007 results in the user seeing information displayed of the type illustrated in *Figure 23*. Each entry, such as entry 2301, includes a check box 2302. Check boxes 2302
25 allow a particular item to be selected by a user such that the actual information file may be supplied to the user from the central database over a communication channel.

What We Claim Is:

1. Apparatus configured to associate files of data of a size greater
5 than a predetermined size, comprising
dividing means configured to divide a file into a plurality of file sections
having a size substantially consistent with a preferred size;
categorising means configured to categorise each of said file sections
to produce sets of section associations; and
10 processing means, configured to process said sets of section
associations to produce a set of category associations for the original
undivided file.
2. Apparatus according to claim 1, wherein said dividing means is
15 configured to divide files into file sections each having a size substantially
consistent with a preferred size, wherein said preferred size is smaller than
said predetermined size.
3. Apparatus according to claim 1 or claim 2, wherein said dividing
20 means is configured to remove tables from a data file before dividing said file
into sections.
4. Apparatus according to any of claims 1 to 3, wherein said
processing means is configured to determine the desirability to increase size
25 sections, and, if such a determination is made, said dividing means is
instructed to increase size sections and repeat the dividing process.

5. Apparatus according to any of claims 1 to 4, including data sources arranged to continually supply data for association.

5 6. Apparatus according to any of claims 1 to 5, wherein said categorising means is configured to categorise each file section by processing a section in combination with association files.

7. Apparatus according to claim 6, including storage means for storing said association files as outline files.

10

8. Apparatus according to claim 7, wherein each of said stored outline file relates to a respective category.

9. Apparatus according to claims 1 to 8, including searching means configured to search categories in response to a user request.

15

10. Apparatus according to claim 9, including output means configured to supply output information identifying files selected by said search.

20

11. A method of associating files of data of a size greater than a predetermined size, comprising steps of

dividing a file into a plurality of file sections each having a size substantially consistent with a preferred size;

25 categorising each of said file sections to produce sets of section associations; and

processing said sets of section associations to produce a set of

category associations for the original undivided file.

12. A method according to claim 11, wherein said preferred size is smaller than said predetermined size.

5

13. A method according to claim 11 or claim 12, wherein tables are removed from a data file before said file is divided into sections.

10 14. A method according to any of claims 11 to 13, wherein an assessment is made as to the desirability to increase size sections, whereafter the size of said section are increased and the dividing process is repeated.

15 15. A method according to claims 11 to 14, wherein data files are continually received from data sources.

20 16. A method according to claims 11 to 15, wherein said categorising is performed by processing a data file in combination with association files.

20

17. A method according to claim 16, wherein said association files are stored as outline files.

25 18. A method according to claim 16, wherein each association file relates to a respective category.

19. A method according to any of claims 11 to 18, wherein

categories are searched in response to a user request.

20. A method according to claim 19, wherein information identifying files is generated in response to said search and returned to a requesting user.

21. A computer system programmed to execute stored instructions such that in response to said stored instructions said system is configured to:

divide a file into a plurality of file sections having a size substantially consistent with a preferred size;

categorise each of said file sections to produce sets of section associations; and

process said sets of section associations to produce a set of category associations for the original undivided file.

22. A computer system programmed to execute stored instructions according to claim 21, configured to divide files into file sections each having a size substantially consistent with a preferred size, wherein said preferred size is smaller than said predetermined size.

23. A computer system programmed to execute stored instructions according to claim 21 or claim 22, configured to remove tables from a data file before dividing said file into sections.

24. A computer system programmed to execute stored instructions according to any of claims 21 to 23, configured to determine the desirability to increase size sections and, if such a determination is made, to increase size

sections and to repeat the dividing process.

5 **25.** A computer system programmed to execute stored instructions according to claim 21 to 24, configured to continually supply data for association.

10 **26.** A computer system programmed to execute stored instructions according to claim 21 to 25, configured to categories each file section by processing a section in combination with association files.

15 **27.** A computer system programmed to execute stored instructions according to claim 26, configured to store said association files as outline files.

20 **28.** A computer system programmed to execute stored instructions according to claim 27, configured to store an outline file relating to each category for which association is required.

25 **29.** A computer system programmed to execute stored instructions according to any of claims 21 to 28, configured to search categories in response to a user request.

30 **30.** A computer system programmed to execute stored instructions according to claim 29, configured to supply output information identifying files selected by said search.

31. A computer-readable medium having computer-readable

instructions executable by a computer such that, when executing said instructions, a computer will perform the steps of:

dividing a file into a plurality of file sections each having a size substantially consistent with a preferred size;

5 categorising each of said file sections to produce sets of section associations; and

processing said sets of section associations to produce a set of category associations for the original undivided file.

10 **32.** A computer-readable medium having computer-readable instructions according to claim 31, such that when executing said instructions a computer also performs the step of removing tables from a data file before dividing the file into sections.

15 **33.** A computer-readable medium having computer-readable instructions according to claim 31 or claim 32, such that when executing said instructions a computer will also perform the steps of assessing the desirability to increase the size of file sections and, where appropriate, increasing the size of said sections and repeating the file division process.

20 **34.** A computer-readable medium having computer-readable instructions according to any of claims 31 to 33, such that when executing said instructions a computer will also perform the step of categorising by processing a data file in combination with association files, preferably stored
25 in the form of outline files.

35. A computer-readable medium having computer-readable

instructions according to any of claims 31 to 34, such that when executing said instructions a computer will also perform the step of searching categories in response to user requests.

- 5 **36.** A computer-readable medium having computer-readable instructions according to claim 35, such that when executing said instructions a computer will also perform the step of generating information identifying files in response to said search and returning said identifying information to a requesting user.

1/23

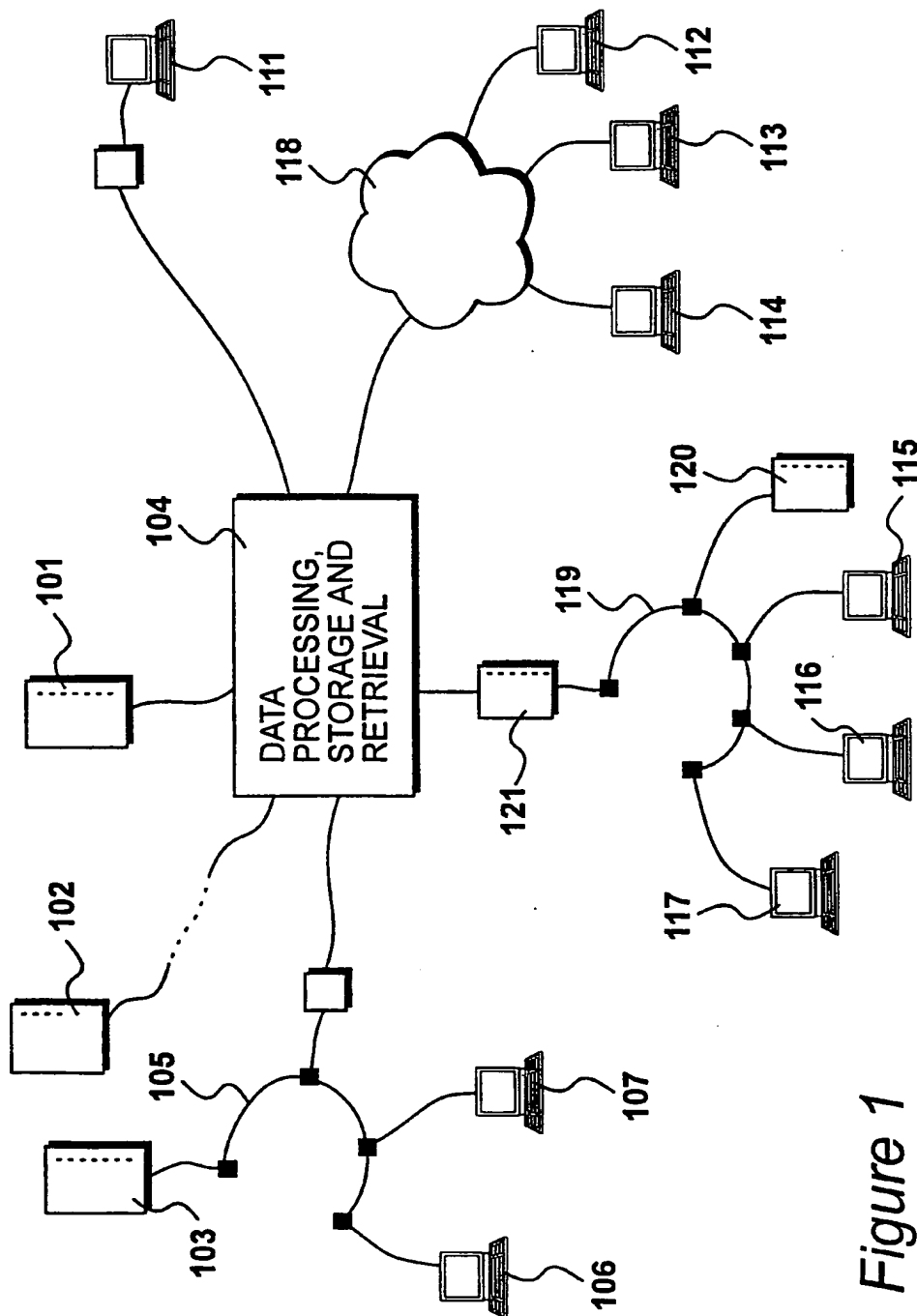
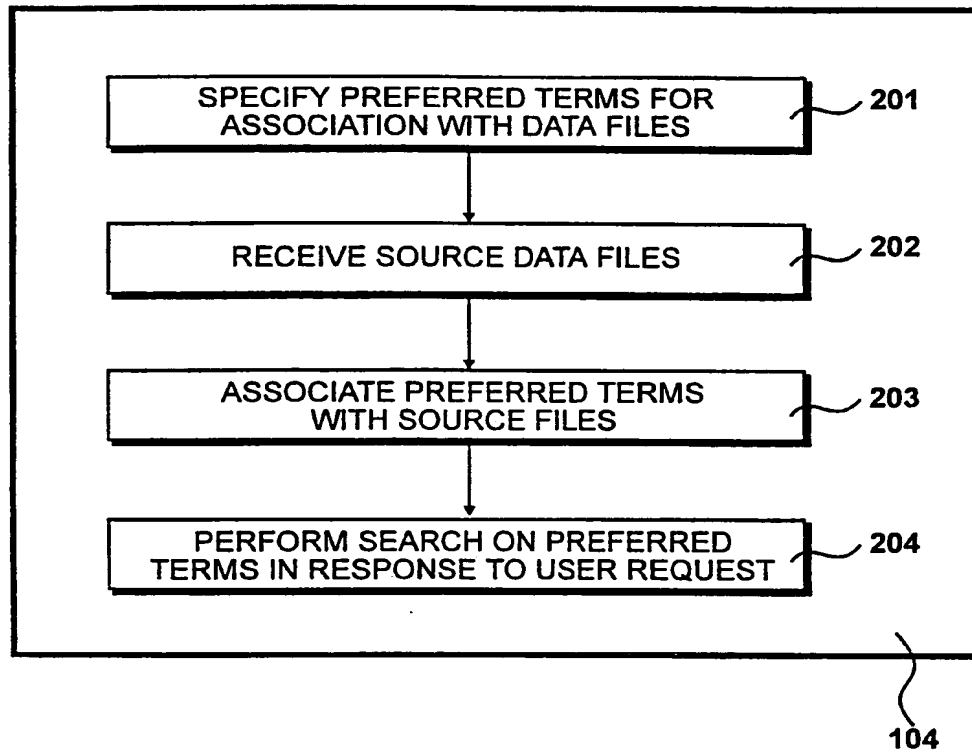


Figure 1

2/23

*Figure 2A*

3/23

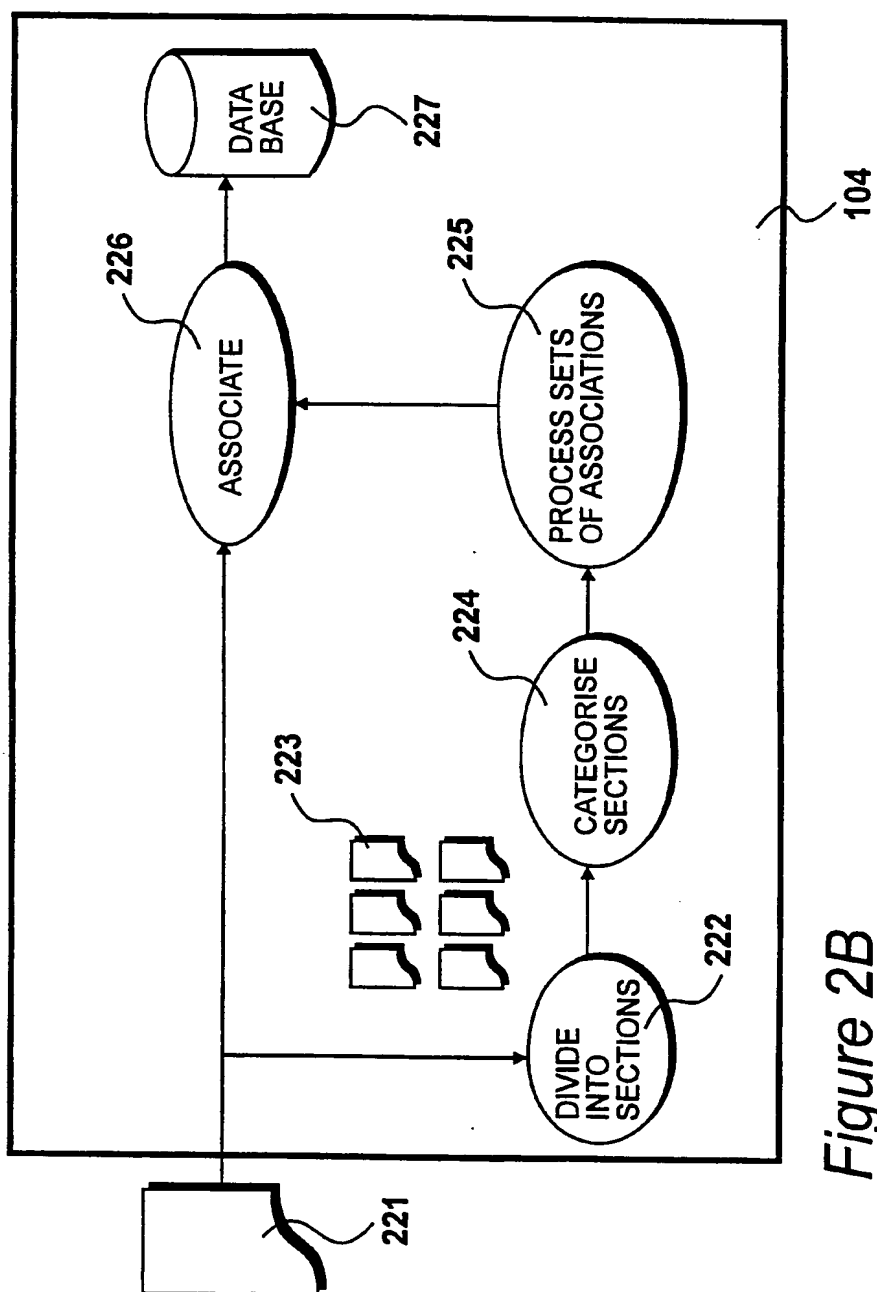


Figure 2B

4/23

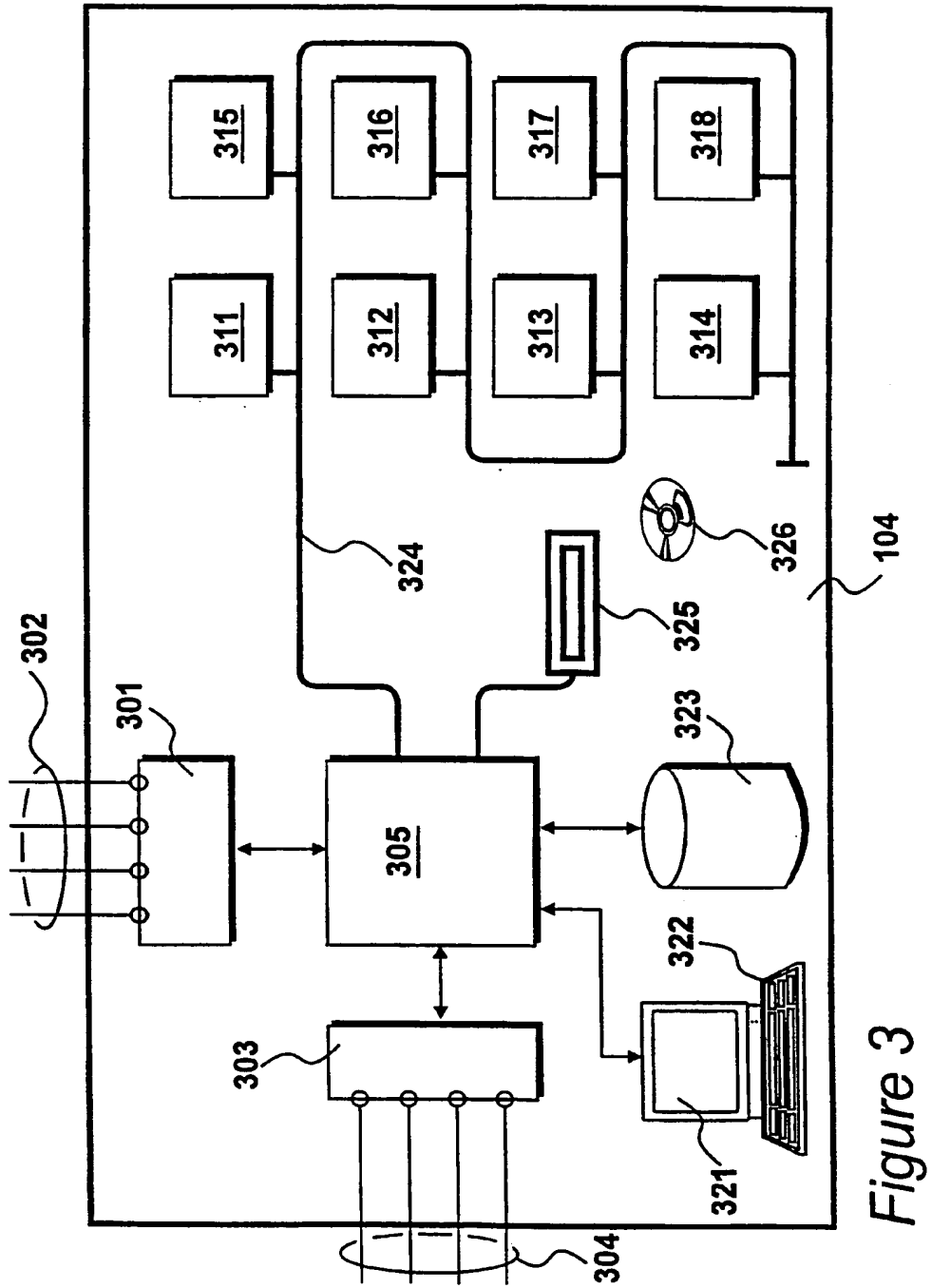
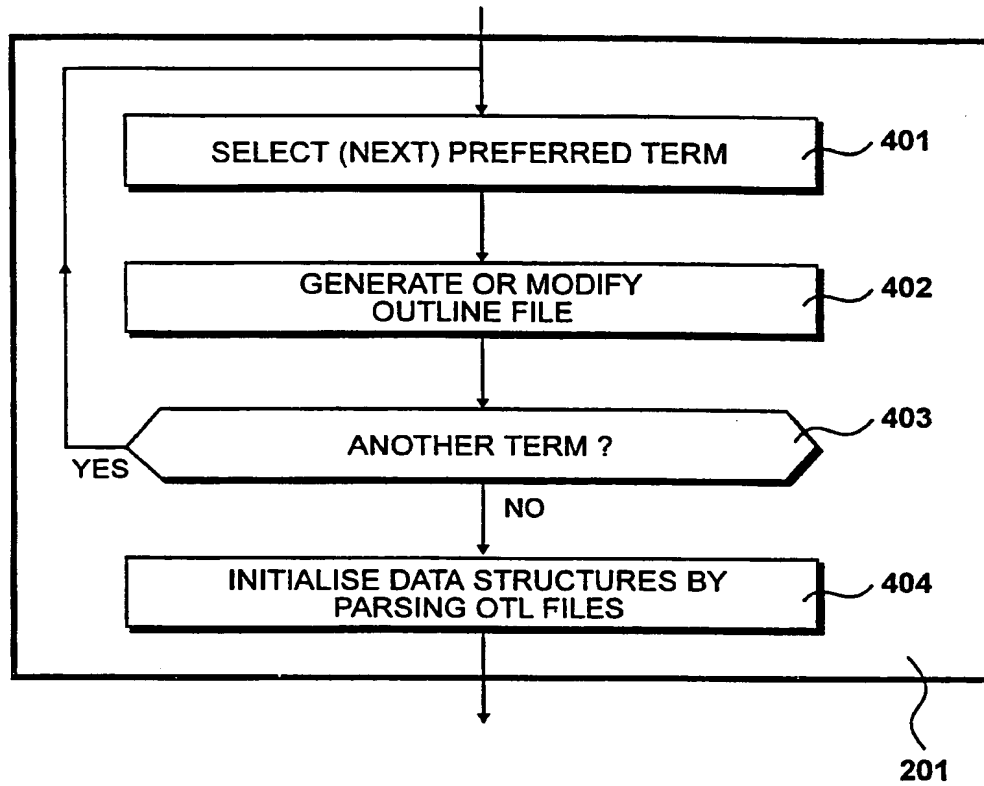
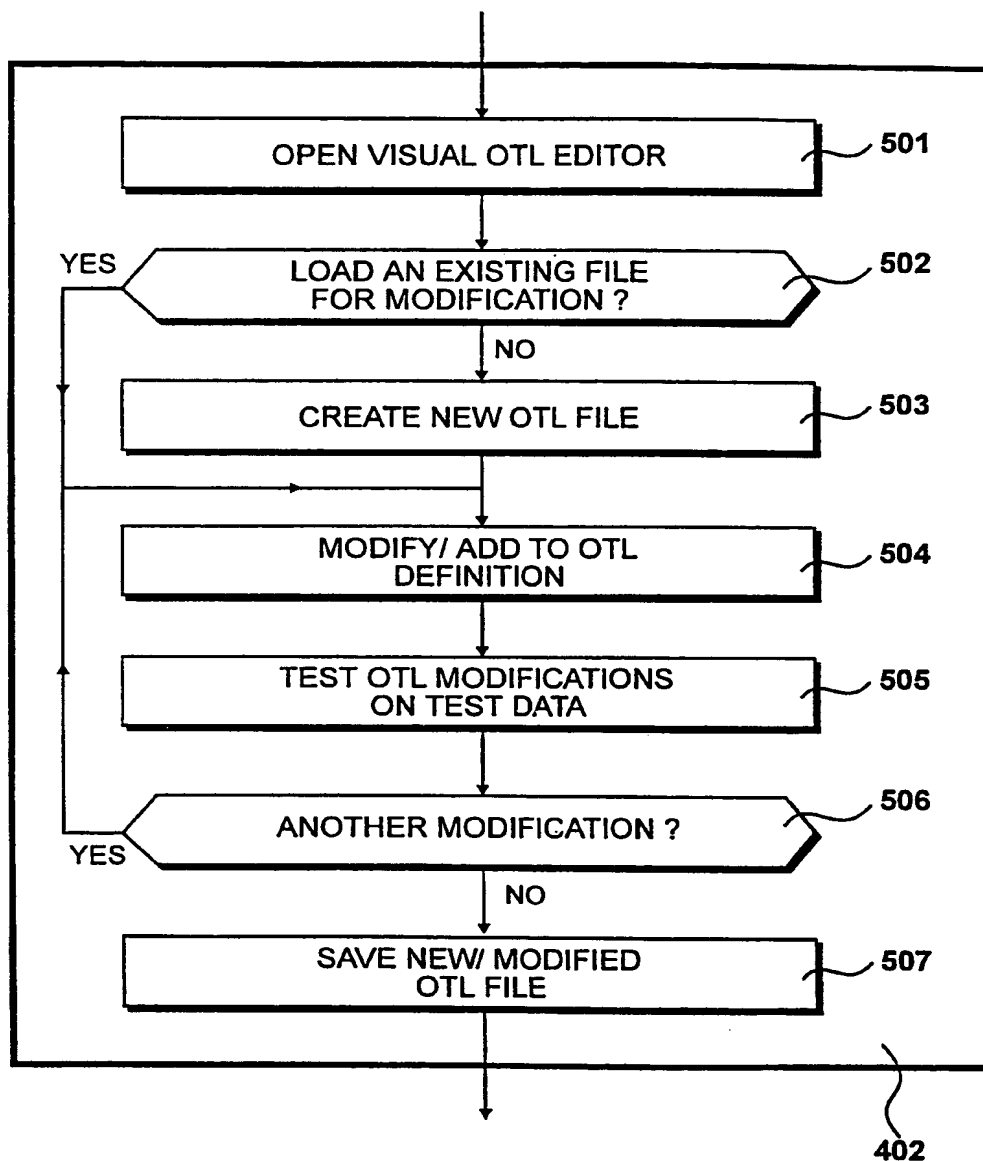


Figure 3

5/23

*Figure 4*

6/23

*Figure 5*

7/23

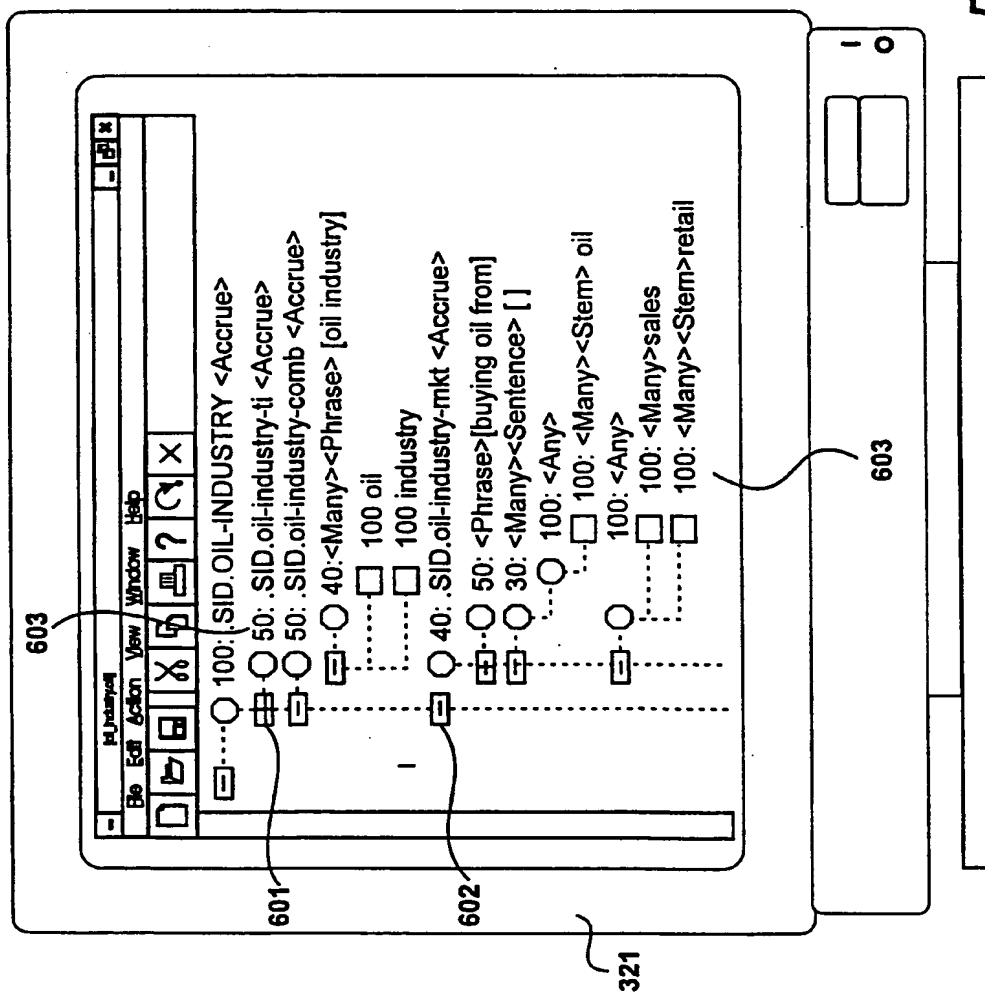


Figure 6

8/23

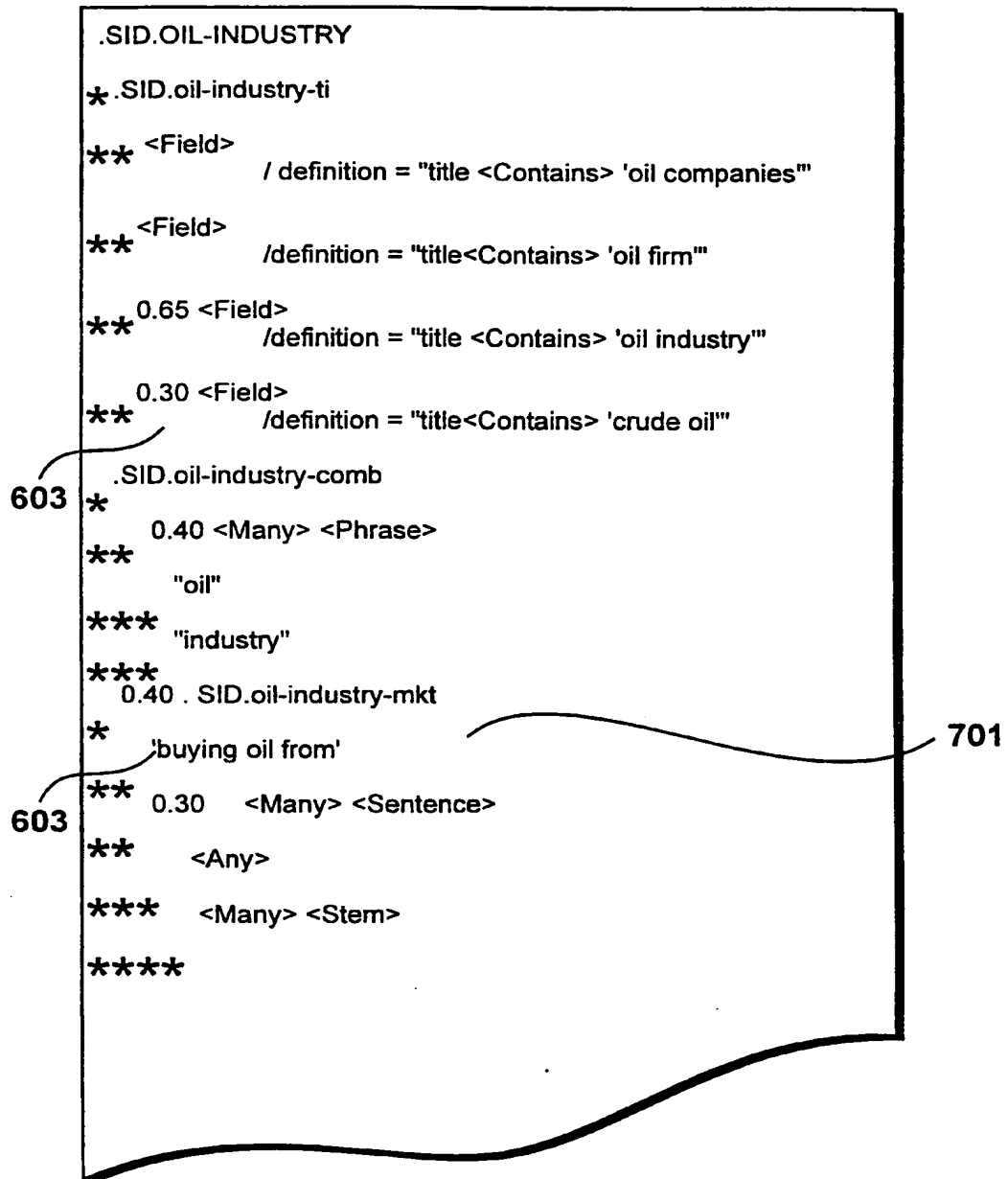


Figure 7

9/23

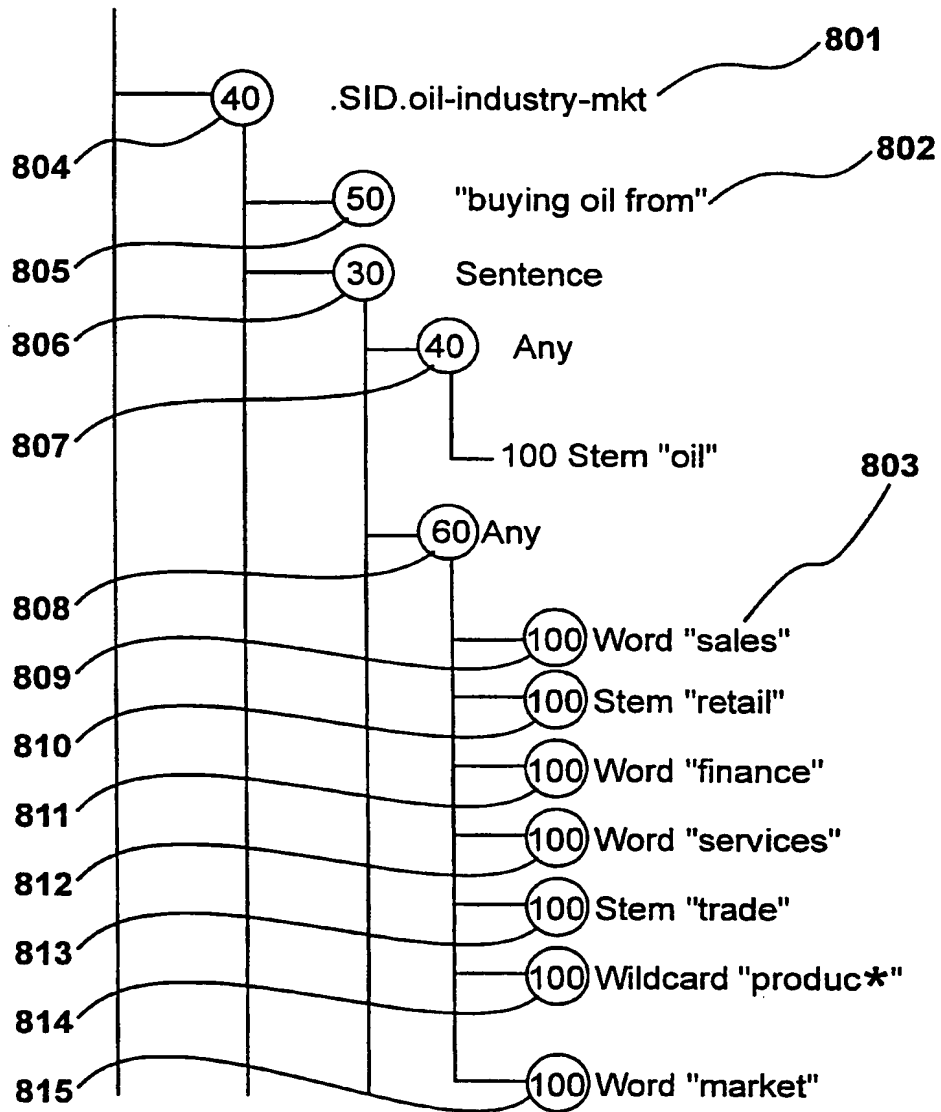
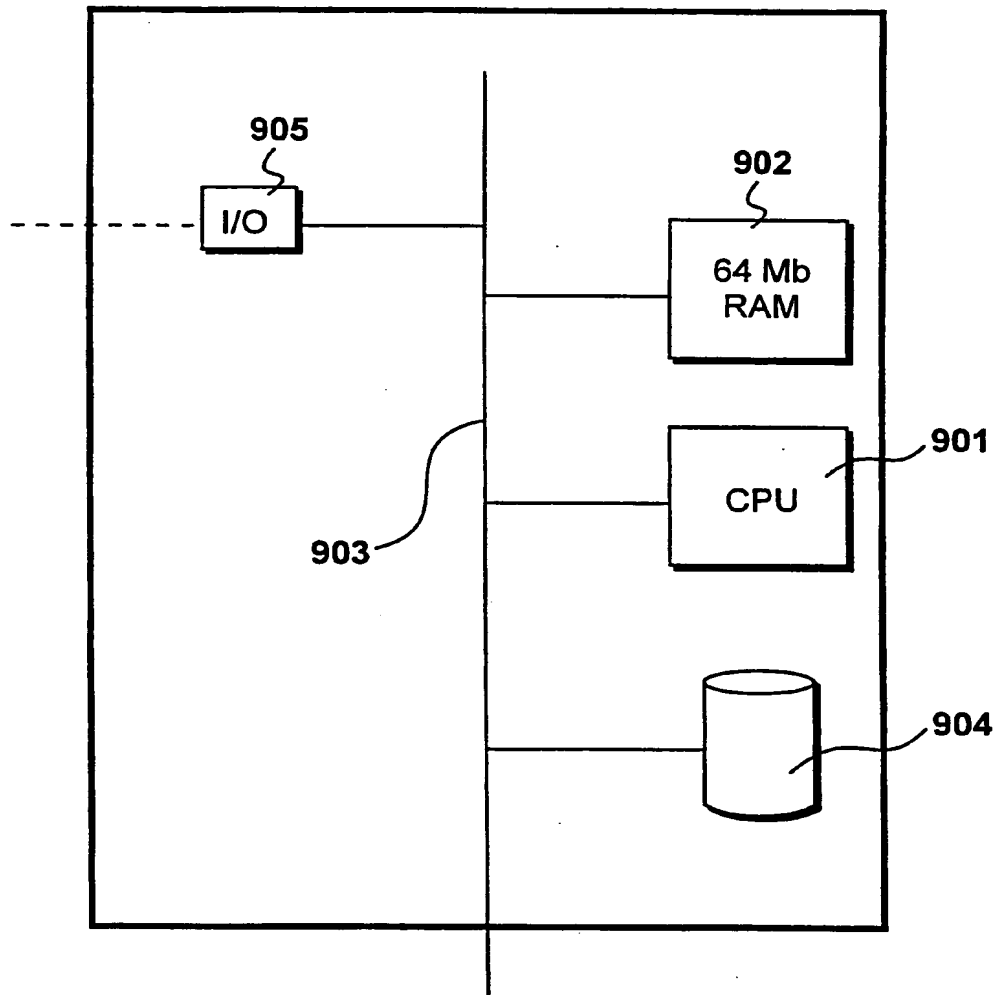


Figure 8

10/23

*Figure 9*

11/23

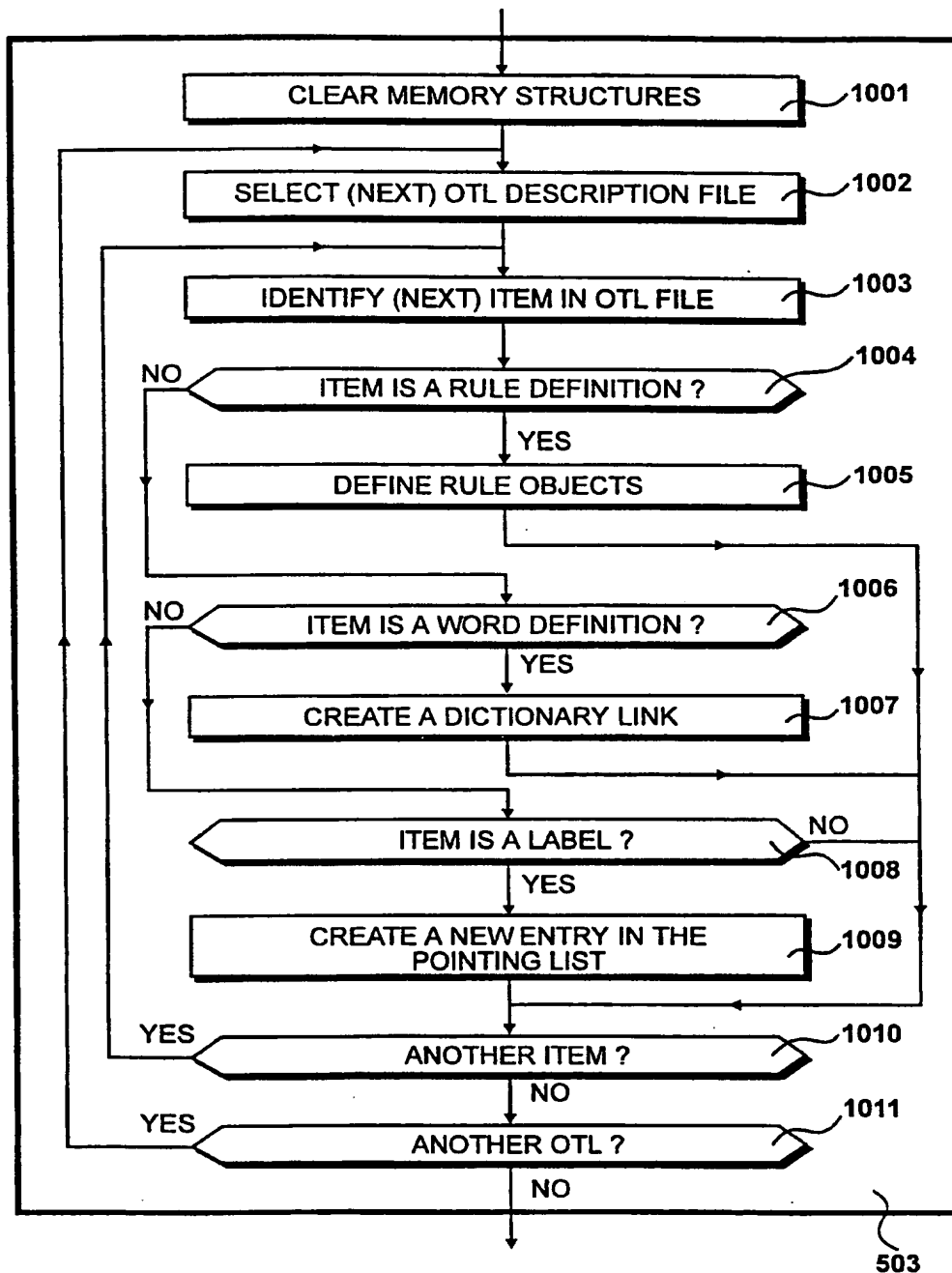


Figure 10

12/23

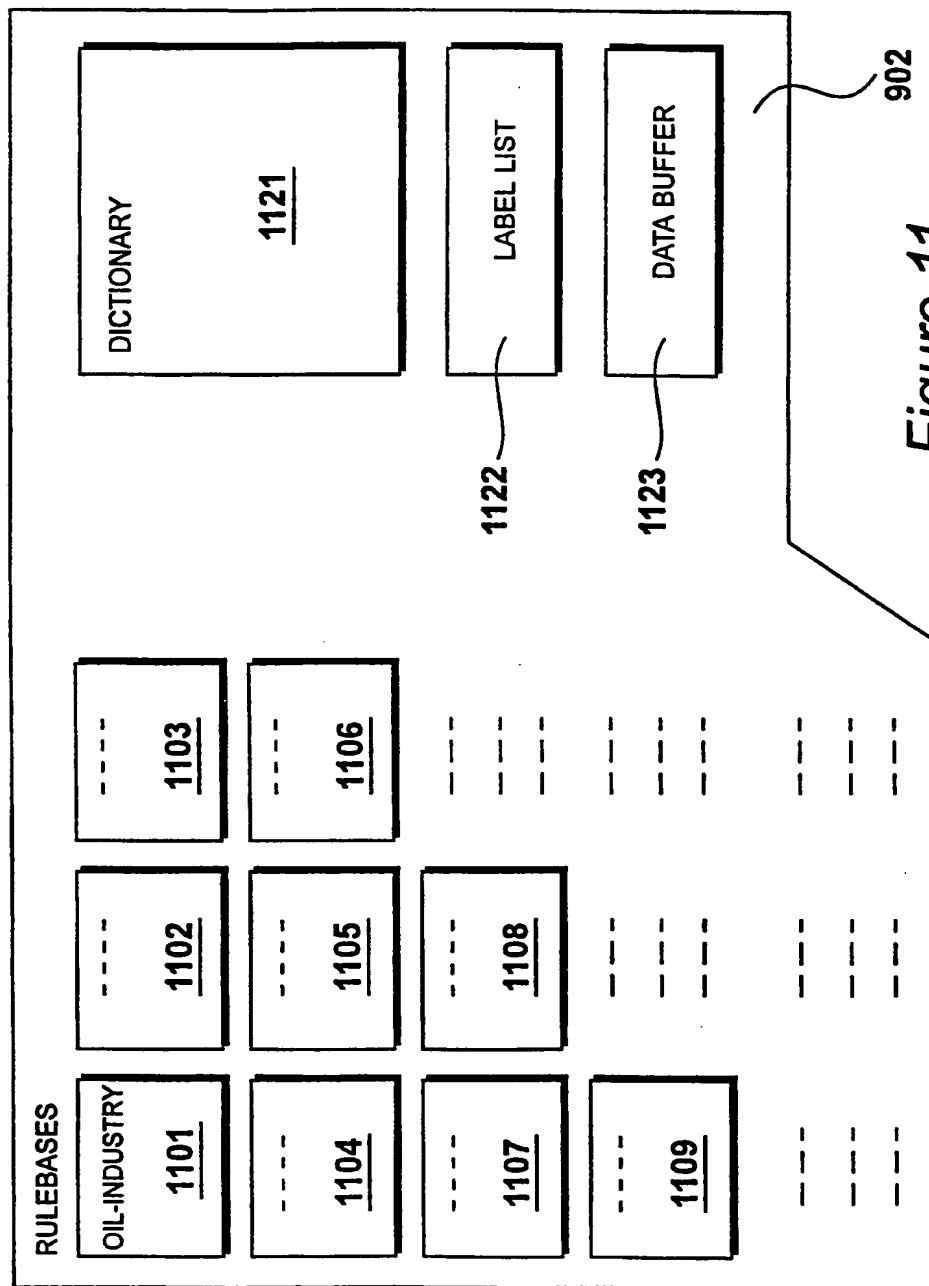


Figure 11

13/23

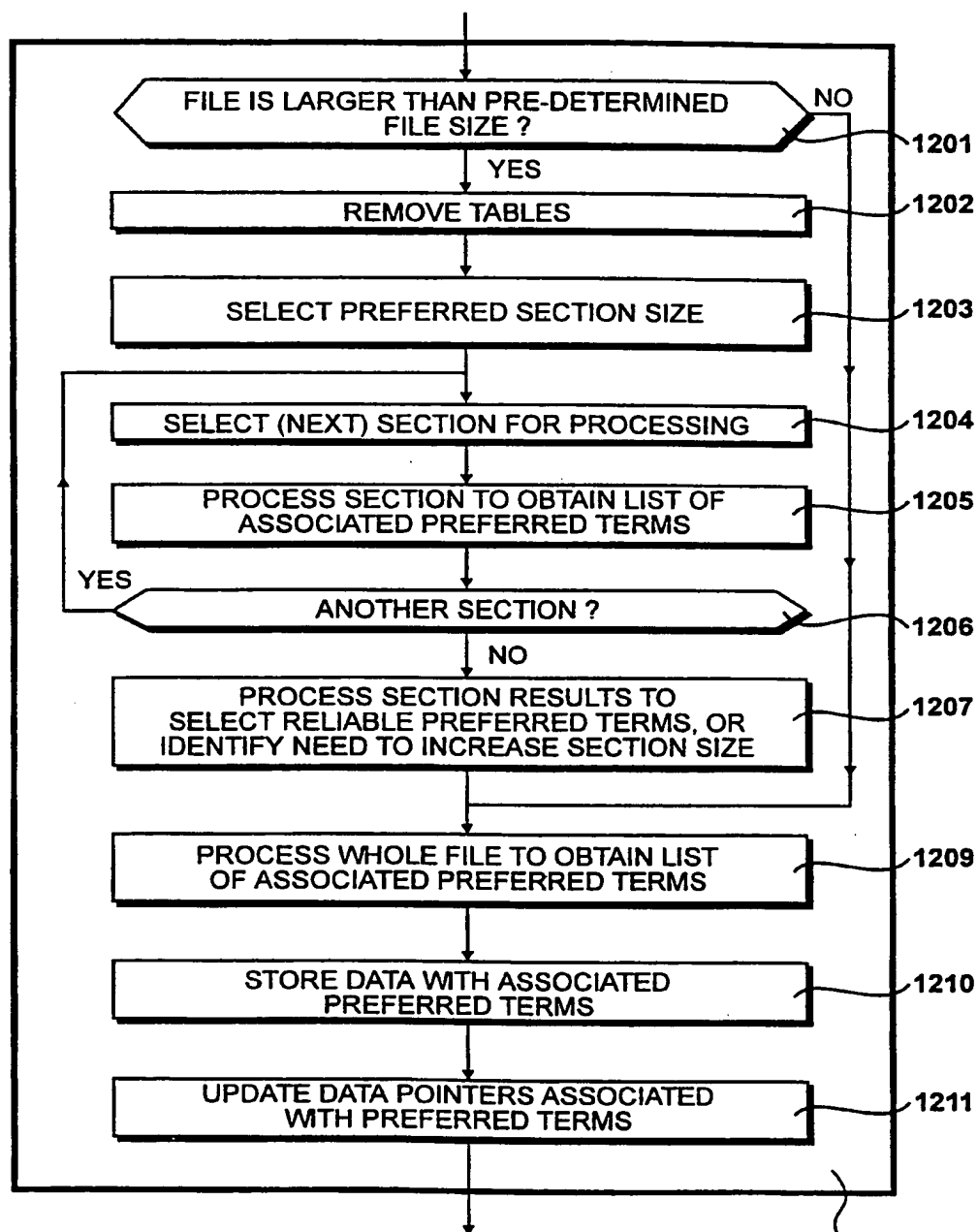
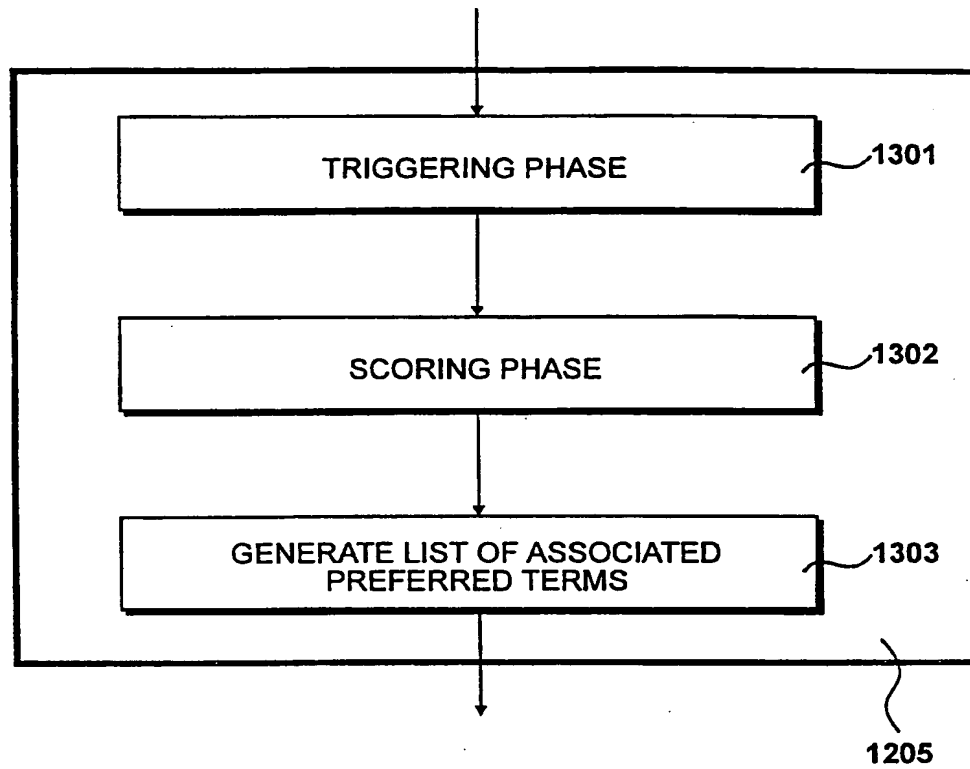


Figure 12

203

14/23

*Figure 13*

15/23

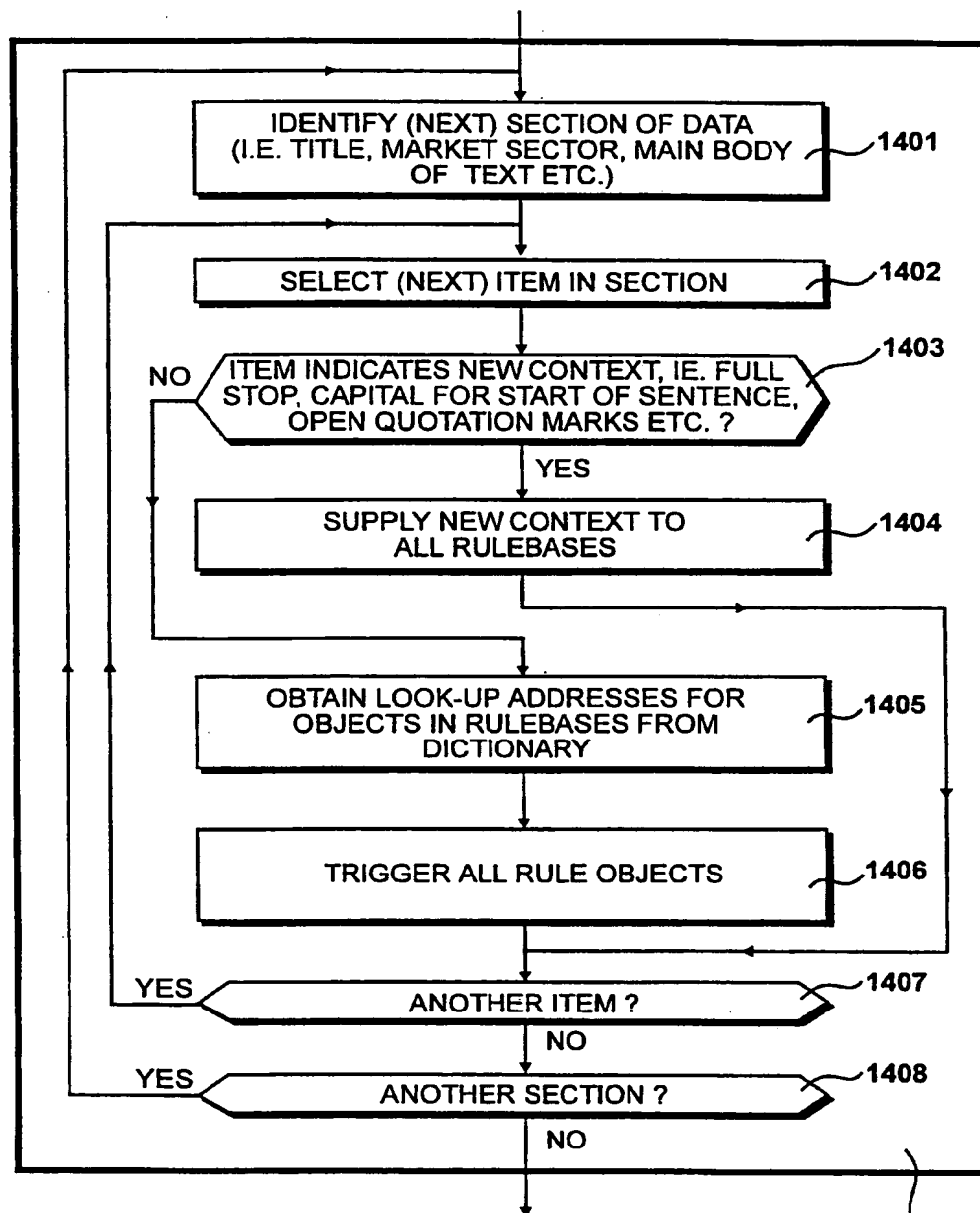
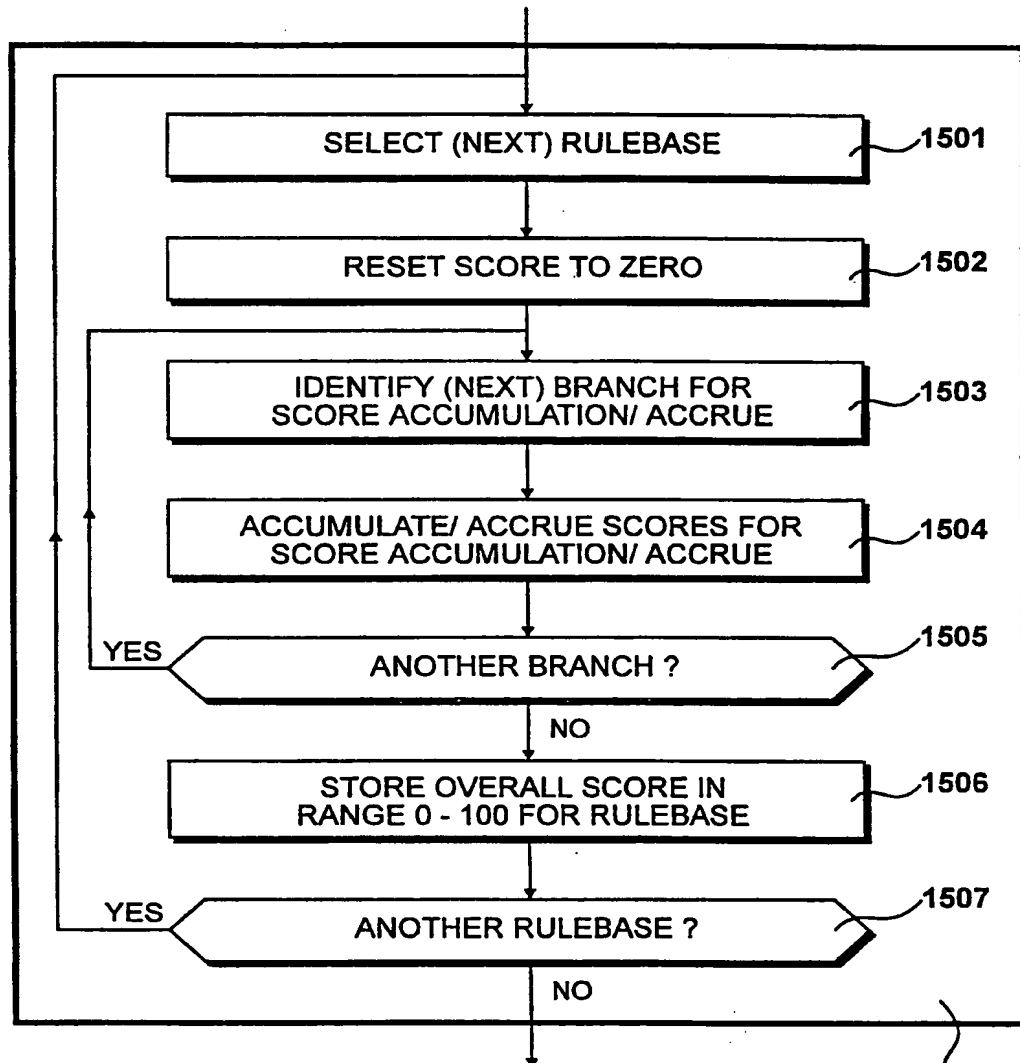


Figure 14

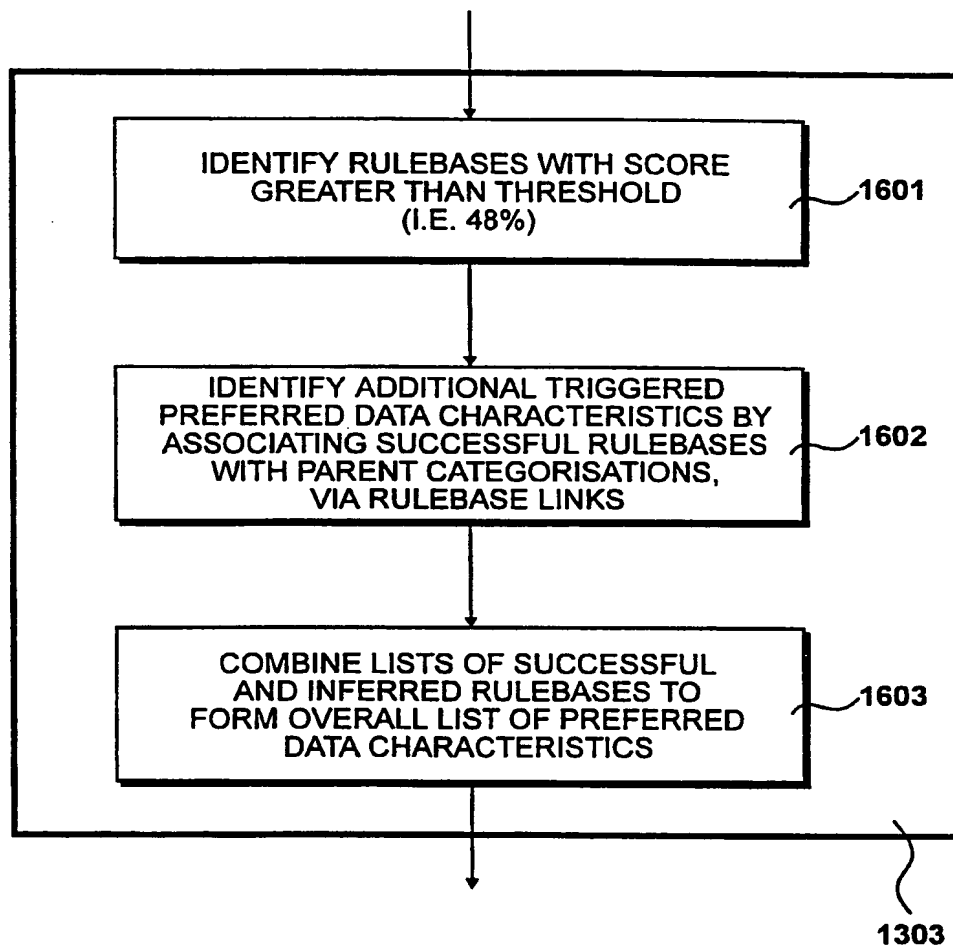
1301

16/23

*Figure 15*

1302

17/23

*Figure 16*

18/23

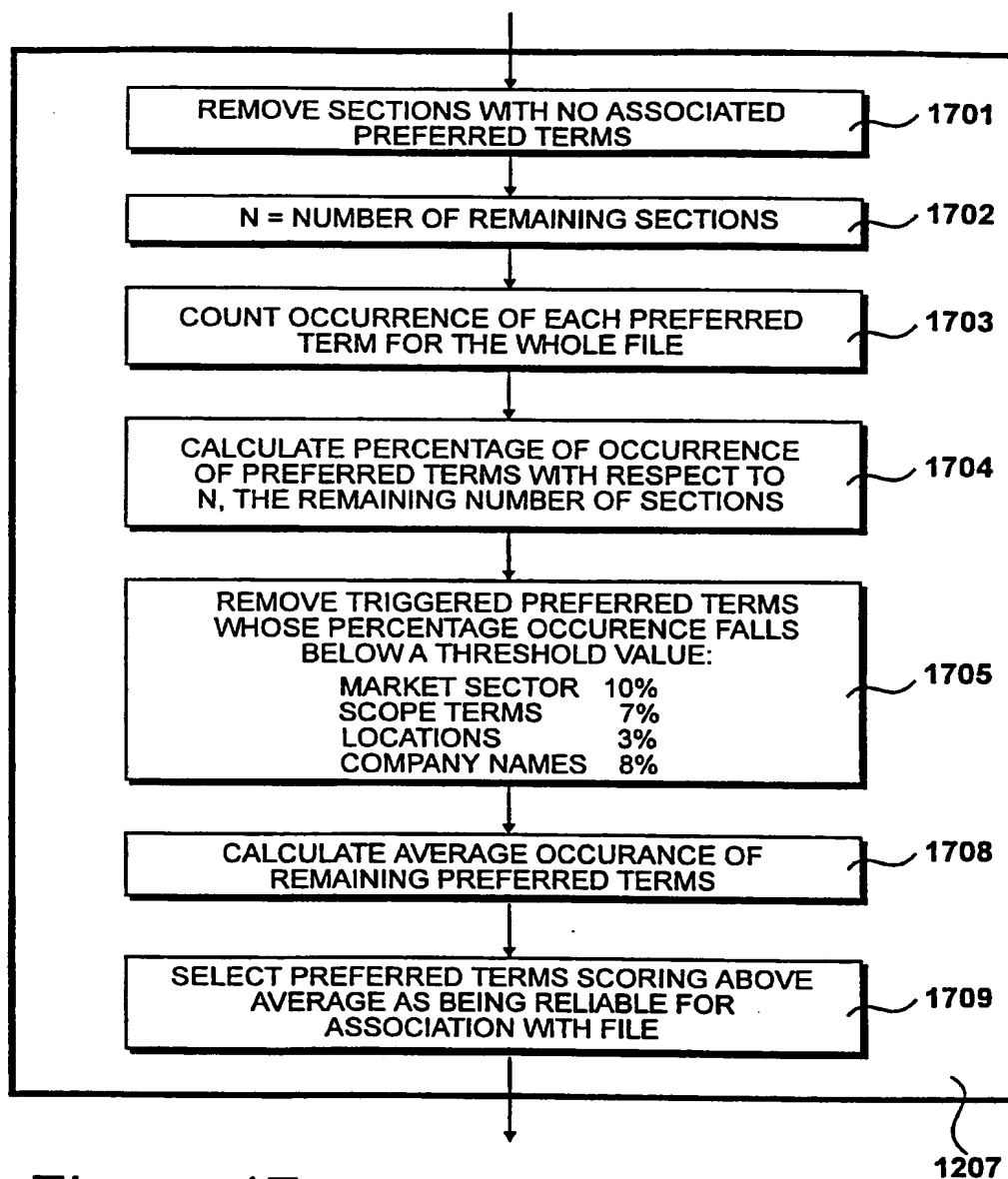


Figure 17

19/23

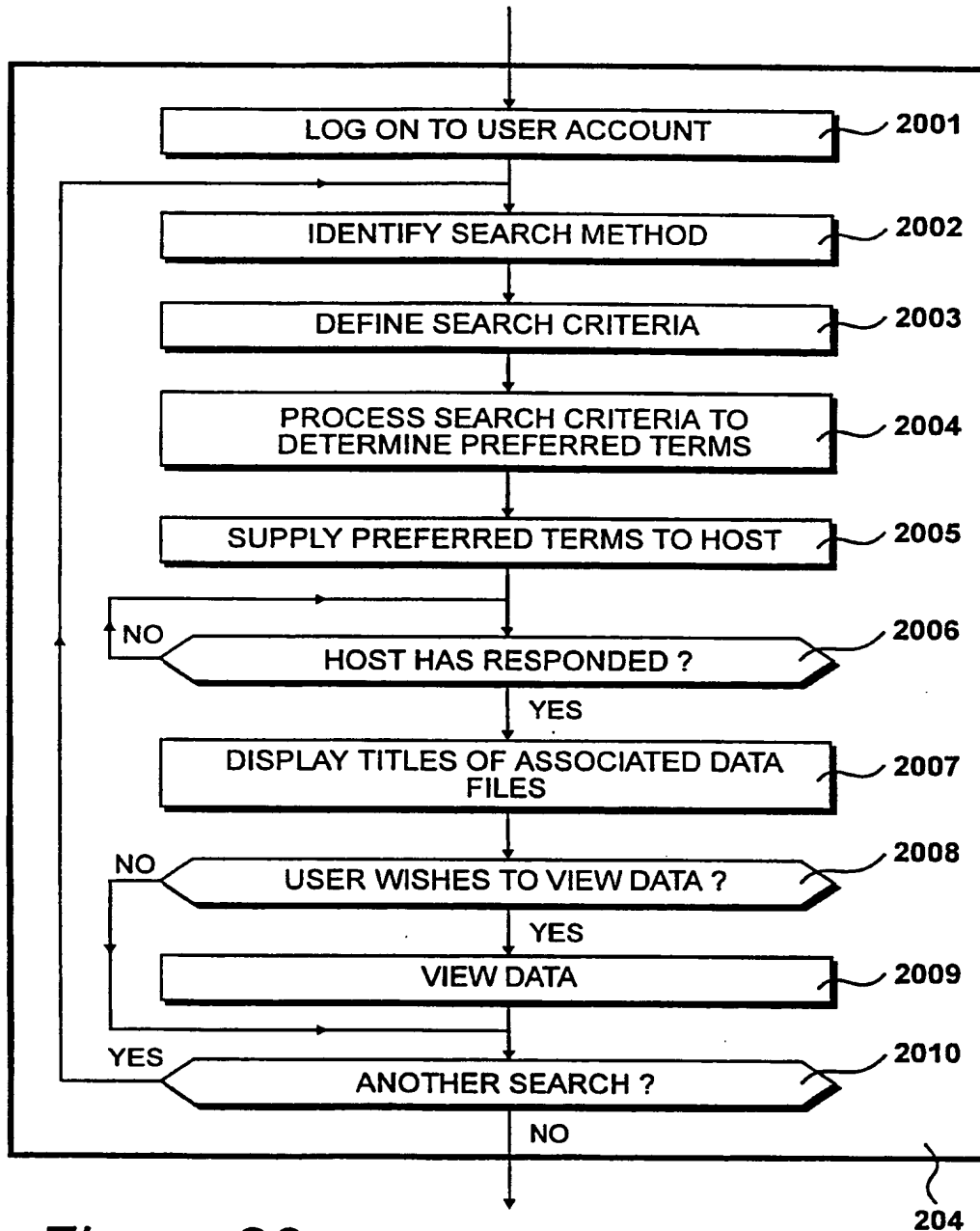
1801 PREFERRED TERM	1802 POINTER
OIL INDUSTRY	OF8912
OIL_INSTITUTIONS	192AC3
OIL_	516321
PETROLEUM_	3200FI
⋮	⋮
⋮	⋮
⋮	⋮
⋮	⋮
⋮	⋮
⋮	⋮

Figure 18

1901 ADDRESS	1902 FILE NAME	1903 POINTER
OF8912	Oil_industry_netherland_3	OF8A20
OF8A20	Oil_ind_india_flash_	OF8193
OFA193	Petrochem_times.3.9.97	100AB1
100AB1	[END]	000000
⋮		
192AC3	BP.index_ft_uk_97	20A21B
⋮		
⋮		
⋮		

Figure 19

20/23

*Figure 20*

21/23

The image shows a web application interface with a navigation bar at the top containing links: [home](#), [Database](#), [News](#), [Titles](#), [Sort](#), [Pub. Date](#), and [Ascending](#). Below the navigation bar, there are several search filters and options:

- Market Sector:** A text input field.
- Companies:** A text input field.
- Free text: stage:** A text input field.
- Title:** A text input field.
- Pub. Date:** A dropdown menu with 'None' selected.
- Countries:** A text input field.
- Publisher:** A text input field.
- Scope:** A text input field.
- From:** A text input field.
- To:** A text input field.
- Ascending:** A checkbox.
- Use Saved Search:** A button.

At the bottom of the interface, there is a footer with links: [Home](#), [Dossier](#), [Portfolio](#), [Alert Manager](#), [Utilities](#), [Client Resources](#), and [Help?](#).

Figure 21

22/23

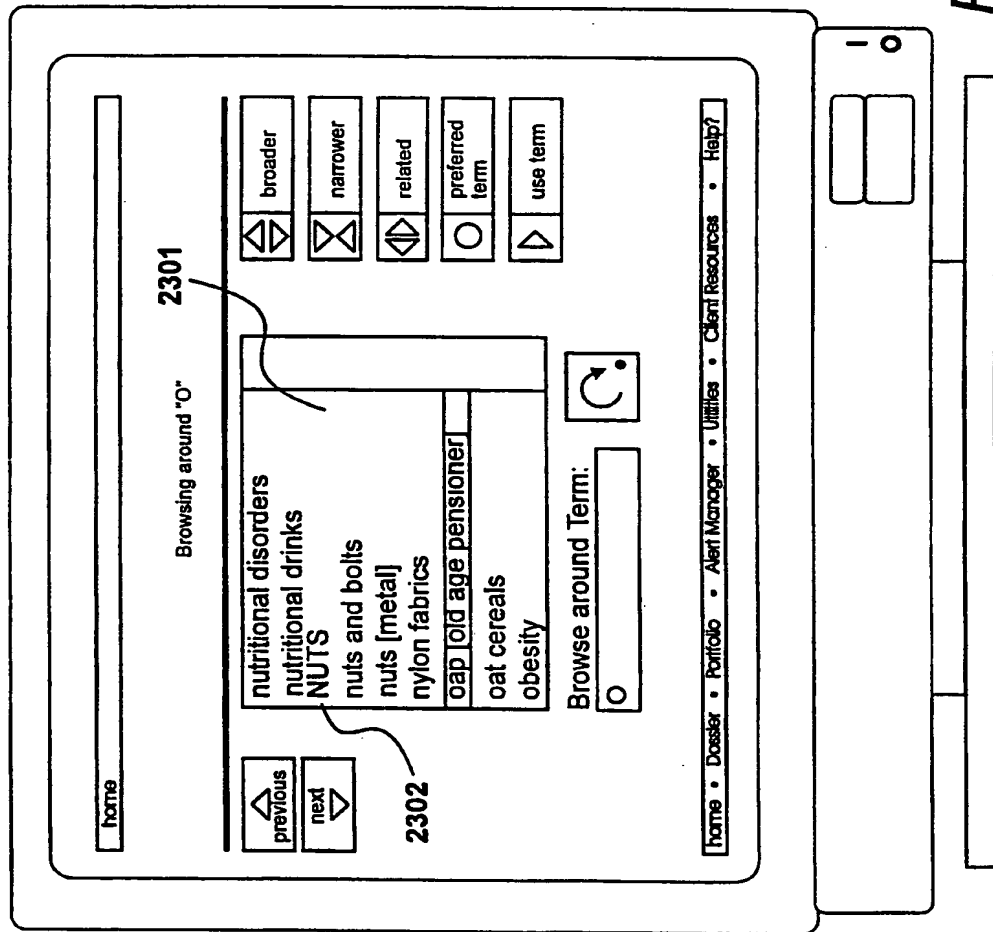


Figure 22

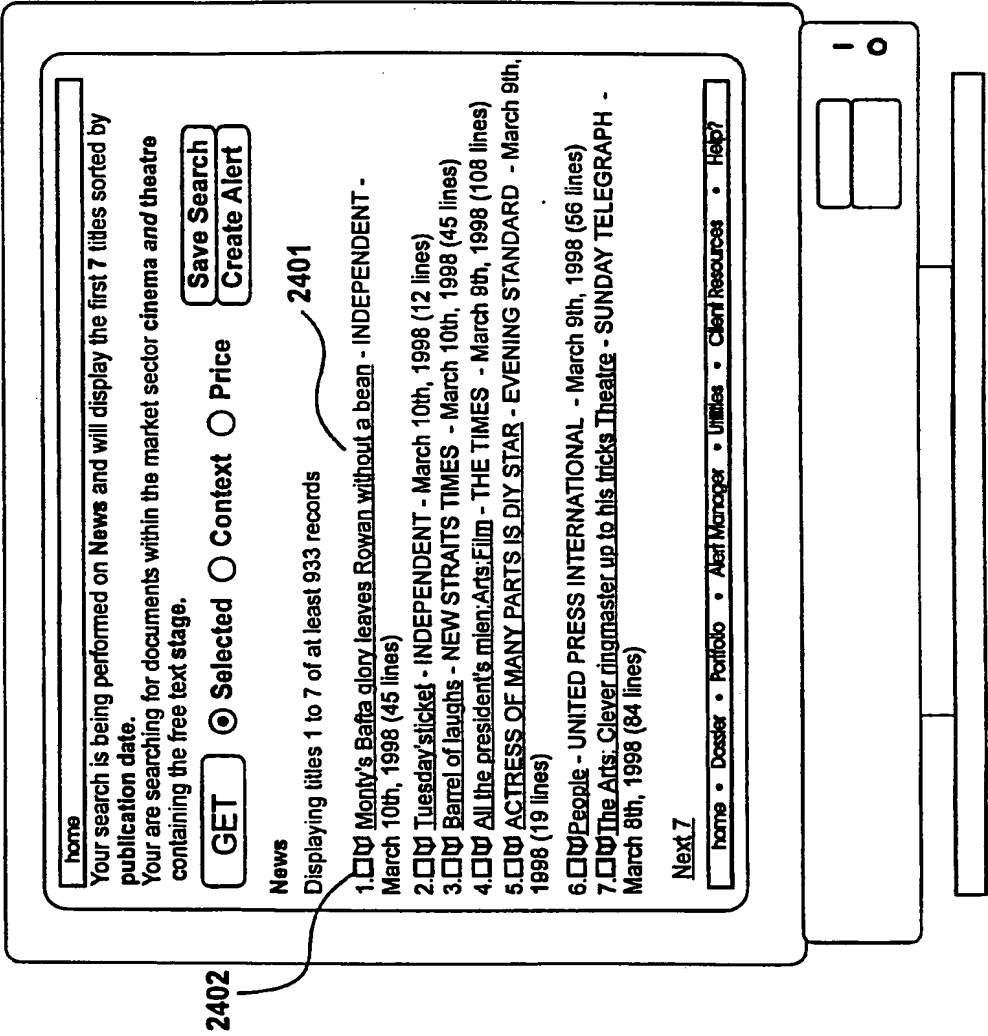


Figure 23

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 99/01213

A. CLASSIFICATION OF SUBJECT MATTER IPC 6 G06F17/30		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) IPC 6 G06F		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practical, search terms used)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	HEARST M A ET AL: "Subtopic structuring for full-length document access" SIXTEENTH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, PITTSBURGH, PA, USA, 27 JUNE-1 JULY 1993, vol. spec. issue., pages 59-68, XP002111439 SIGIR Forum, 1993, USA ISSN: 0163-5840 abstract page 3, left-hand column, line 39 - page 4, right-hand column, line 30 page 9, left-hand column, paragraph 5 - page 10 <div style="text-align: center;">--- -/--</div>	1,2, 4-12, 14-31, 33-36
<div style="display: flex; justify-content: space-between;"> <input checked="" type="checkbox"/> Further documents are listed in the continuation of box C. <input type="checkbox"/> Patent family members are listed in annex. </div>		
* Special categories of cited documents :		
<div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p> </div> <div style="width: 45%;"> <p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.</p> <p>"&" document member of the same patent family</p> </div> </div>		
Date of the actual completion of the international search <div style="text-align: center;">5 August 1999</div>		Date of mailing of the international search report <div style="text-align: center;">17/08/1999</div>
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016		Authorized officer <div style="text-align: center;">Fournier, C</div>

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 99/01213

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No. -
Y	HAYES P J: "TCS: A SHELL FOR CONTENT-BASED TEXT CATEGORIZATION" PROCEEDINGS OF THE CONFERENCE ON ARTIFICIAL INTELLIGENCE APPLICATION, SANTA BARBARA, MAR. 5 - 9, 1990, vol. 1, no. CONF. 6, 5 March 1990 (1990-03-05), pages 320-326, XP000295098 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS ISBN: 0-8186-2032-3 the whole document	1,2, 4-12, 14-31, 33-36
A	PATENT ABSTRACTS OF JAPAN vol. 018, no. 287 (P-1746), 31 May 1994 (1994-05-31) & JP 06 052162 A (FUJITSU LTD), 25 February 1994 (1994-02-25) abstract	1,11,21, 31
A	"INTELLIGENT LIBRARY FILTER FOR OFFICE" IBM TECHNICAL DISCLOSURE BULLETIN, vol. 34, no. 4B, 1 September 1991 (1991-09-01), pages 32-33, XP000189538 ISSN: 0018-8689 the whole document	1,5, 9-11,15, 19-21, 25, 29-31, 35,36

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/GB 99/01213

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
JP 06052162 A	25-02-1994	NONE	